Appeal Brief under 37 C.F.R. § 41.37	Attorney Docket No.	3035-101
	First Named Inventor	**Karen Uhlmann**
	Title: **Method of Detecting Epigenetic Biomarkers by Quantitative methylSNP Analysis**	
	Application No.	**10/823,784**
	371(c) Date	April 14, 2004
	Group Art Unit	1634
	Examiner Name	**Amanda Marie Shaw**

MAIL STOP APPEAL BRIEF
Commissioner for Patents
P.O. Box 1450
Alexandria, VA 22313-1450

On July 23, 2009 the appellants appealed from the final rejection of claim 1 to 5, 7 to 20 and 22 to 39. For the consideration of this appeal brief, a four months extension of time is submitted herewith.

It will be shown below that these rejections are untenable and that the appealed claims 1 to 5, 7 to 20 and 22 to 39 patentably distinguish the appellants' invention over the references applied there against.

The Commissioner is authorized to charge any deficiency of payment to Deposit Account No. 50-3135.

TABLE OF CONTENTS

TABLE OF AUTHORITIES

REAL PARTY IN INTEREST

The real party in interest in this appeal is Max-Delbrück-Centrum für Molekulare Medizin, the assignee of record and having its principal office at Robert-Roessle-Strasse 10, 13125 Berlin, Germany.

The assignee acquired its interest by virtue of assignment and recorded by the assignment branch of the United States Patent and Trademark Office at reel 016624, frame 0483.

Nonwithstanding the assignment, the appellants, Karen Uhlmann, Peter Nürnberg, and Anja Brinckmann remain a party of interest because of the German law (Gesetz über Arbeitnehmererfindungen) covering employed inventors.

RELATED APPEALS AND INTERFERENCES

Appellants' attorney is unaware of any other appeals or interferences which will directly affect or be directly affected by or have a bearing on the Board's decision in the pending appeal.

STATUS OF CLAIMS

Rejected Claims:	1 to 5, 7 to 20 and 22 to 39
Allowed Claims:	None
Withdrawn Claims:	None
Claims Objected to:	None
Claims Cancelled:	6 and 21
The appealed claims are:	1 to 5, 7 to 20 and 22 to 39

Claims 1 to 5, 7 to 20 and 22 to 39 are pending in the application. Claims 6 and 21 were cancelled. Claims 1 to 5, 7 to 20 and 22 to 39 stand twice rejected. Since, as of the date of this appeal, none of pending claims 1 to 5, 7 to 20 and 22 to 39 have been allowed, this appeal brief follows.

STATUS OF AMENDMENTS

Claims 1 to 5, 7 to 20 and 22 to 39 were rejected in the final action mailed on January 23, 2009. No subsequent amendments were filed.

SUMMARY OF CLAIMED SUBJECT MATTER

The claims cover a method for detecting the methylation status of a nucleotide of a nucleic acid molecule. The method comprises treating a sample comprising the nucleic acid molecule with an agent that converts said nucleotide when it is methylated form or non-methylated form so that it pairs with a nucleotide with which it would normally not pair with. The so treated nucleic acid molecule is then amplified with an amplification primer that is detectably labelled with a detectable label that forms an anchor for removal of the single stranded nucleic acid molecules. The single stranded nucleic acid molecule so generated is then real-time sequenced and the methylation status of the nucleotide in the sample is detected or determined.

The method combines the treatment of the nucleic acid molecules to create new pairing partners upon subsequent amplification, amplification and real-time sequence to provide a highly efficient method to detect, and optionally quantify, the methylation status of nucleotides in a nucleic acid molecule that is amenable to high throughput analysis, e.g., on microtiter plates which allows, e.g., 96 different gene loci may be screened. Pathological condition or the predisposition for said pathological condition may be diagnosed using the method.

Claim 1

The appellant's invention in independent claim 1 is directed to a method for detecting the methylation status of a nucleotide at a predetermined position in a nucleic acid molecule (page 4, lines 19 to 20; page 8, lines 17 to 20). The methylation status is detected in a sample comprising the nucleic acid molecule (page 4, lines 21 to 22; page 5, lines 23 to 28). The sample is treated in an aqueous solution (page 4, line 22; page 6, lines 20 to 23) with an agent suitable for the conversion of said nucleotide if present in (i) methylated form; or (ii) non-methylated form (page 4, lines 22 to 24, page 6, line 28 to page 7, line 2) to pair with a nucleotide normally not pairing with said nucleotide prior

to conversion (page 4, lines 23 to 24; page 7, lines 2 to 12). The so treated nucleic acid molecule is amplified via at least one amplification primer (page 4, lines 24 to 25; page 13, lines 3 to 6) to produce an amplification product (page 13, lines 6 to 7) and the amplification product is converted into single stranded amplified nucleic acid molecules (page 13, lines 6 to 9, Figure 1 A, Figure 1B, left hand side). The at least one amplification primer is detectably labeled with a detectable label that forms an anchor for removal of said single stranded amplified nucleic acid molecules (page 13, lines 3 to 20, Figure 1B, left hand side) to generate a single stranded amplified nucleic acid molecule (page 13, lines 4 to 9). The single stranded amplified nucleic acid molecule is real-time sequenced (page 4, line 25; page 13, lines 23 to 30) and it is detected whether said nucleotide is methylated or not methylated at said predetermined position in the sample (page 4, 26 to 27; page 14, lines 1 to 5).

Claim 12

Independent claim 12 describes a method for diagnosing a pathological condition or the predisposition for a pathological condition by determining the methylation status of a nucleotide at a predetermined position in the nucleic acid molecule (page 17, line 24 to 27; and in Figures 5 and 6; page 8, lines 17 to 20). The methylation status is detected in a sample comprising the nucleic acid molecule (page 17, line 27 to 28; page 5, lines 23 to 28). The sample is treated in an aqueous solution (page 17, line 28; page 6, lines 20 to 23) with an agent suitable for the conversion of said nucleotide if present in (i) methylated form; or (ii) non-methylated form (page 17, lines 29 to 31; page 6, line 28 to page 7, line 2) to pair with a nucleotide normally not pairing with said nucleotide prior to conversion (page 17, lines 30 to 31; page 7, lines 2 to 12). The so treated nucleic acid molecule is amplified via at least one amplification primer (page 13, lines 3 to 6) to produce an amplification product (page 13, lines 6 to 9) and the amplification product is converted into single stranded amplified nucleic acid molecules (page 13, lines 6 to 9, Figure A, Figure 1B, left hand side). The at least one amplification primer is detectably labeled with a detectable label that forms an anchor for removal of said single stranded amplified nucleic acid molecules (page 13, lines 3 to 20, Figure 1B, left hand side) to generate a single stranded amplified nucleic acid molecule (page 13, lines

5 to 9). The single stranded amplified nucleic acid molecule is real-time sequenced (page 17, line 32; page 13, lines 23 to 30) and it is detected whether said nucleotide is methylated or not methylated at said predetermined position in the sample (page 17, line 33 to page 18, line 3) to diagnose said pathological condition or the predisposition for said pathological condition (page 17, lines 24 to 27).

Claim 32

Independent claim 32 describes a method for generating new nucleotide pairing partners upon amplification of at least one nucleic acid molecule for the detection of the methylation status in a nucleotide sample (page 8, lines 17 to 20).

The methylation status is detected by providing at least one nucleic acid molecule and treating a nucleic acid molecule with an agent suitable for the conversion of said nucleotide if present in methylated form or non-methylated form (page 6, lines 28 to 32) to pair with a nucleotide pairing partners normally not pairing with said nucleotide prior to conversion (page 7, lines 2 to 12). The so treated nucleic acid molecule is amplified via at least one amplification primer (page 11, line 13 to page 13, line 6) to produce an amplification product (page 13, lines 6 to 7) and the amplification product is converted into single stranded amplified nucleic acid molecules (page 13, lines 6 to 9, Figure 1 A, Figure 1B, left hand side). The at least one amplification primer is detectably labeled with a detectable label that forms an anchor for removal of said single stranded amplified nucleic acid molecules (page 13, lines 3 to 20, Figure 1B, left hand side) to generate a single stranded amplified nucleic acid molecule (page 13, lines 4 to 9) comprising said new nucleic acid pairing partners normally not pairing with said nucleotide prior to conversion. The single stranded amplified nucleic acid molecule is real-time sequenced (page 8, line 19; page 13, lines 23 to 30) and determining the amount of said nucleotide pairing with said new nucleotide pairing partners to detect the methylation status of nucleotides of said nucleic acid (page 8, line 10 to 15; page 9, lines 2 to 6).

Claim 10

Claim 10 is dependent from claim 1 (described above) and the method further

comprises calculating a frequency of methylated nucleotides from results of said real-time sequencing (page 3, lines 4 to 5; page 3, lines 10 to 17; Fig. 2, top; Fig. 3, black circles; Fig. 4; page 23, lines 23 to 27).

Claim 25

Claim 25 is dependent from claim 12 (described above) and the method further comprises calculating a frequency of methylated nucleotides from results of said real-time sequencing (page 3, lines 4 to 5; page 3, lines 10 to 17; Fig. 2, top; Fig. 3, black circles; Fig. 4; page 23, lines 23 to 27).

Claim 34

Dependent claim 34 is dependent from claim 10. Claim 10 is dependent from claim 1 (described above) and the method further comprises calculating a frequency of methylated nucleotides from results of said real-time sequencing (page 3, lines 4 to 5; page 3, lines 10 to 17; Fig. 2, top; Fig. 3, black circles; Fig. 4; page 23, lines 23 to 27). Claim 34 further comprises detecting an allele frequency, where an allele frequency of 5% can be detected (Fig. 2 and page 24, lines 7 to 14).

Claim 39

Dependent claim 39 is also dependent from claim 10. Claim 10 is dependent from claim 1 (described above) and the method further comprises calculating a frequency of methylated nucleotides from results of said real-time sequencing (page 3, lines 4 to 5; page 3, lines 10 to 17; Fig. 2, top; Fig. 3, black circles; Fig. 4; page 23, lines 23 to 27). Claim 39 further comprises detecting an allele frequency of 5% with a standard deviation of not more than 1% (Fig. 2 and page 24, lines 7 to 14).

GROUNDS OF REJECTION TO BE REVIEWED ON APPEAL

The grounds of rejection to be reviewed on appeal are as follows.

Whether claims 1-5, 7-9, 11-12, 19-20, 22-24, 26-33 and 36 are obvious under 35 USC §103(a) over Uhlmann et al, CHANGES IN METHYLATION PATTERNS IDENTIFIED BY TWO-DIMENSIONAL DNA FINGERPRINTING, *Electrophoresis* 20: 1748-55 (1999). (hereinafter "Uhlmann '99") in view of U.S. Patent No. 6,258,568 Nyren et al. (hereinafter "Nyren").

Whether claims 12-16, 18 and 38 are obvious under 35 USC §103(a) over Uhlmann '99 in view of Nyren, and further in view of US Patent No. 5,786,146 to Herman (hereinafter "Herman").

Whether claim 17 is obvious under 35 USC §103(a) over Uhlmann '99 in view of Nyren, and Herman as applied to claims 12 and 38, in further view of US Patent Publication No. 2003/0232351 to Feinberg (hereinafter "Feinberg").

Whether claims 10, 25, 34 and 39 are obvious under 35 USC §103(a) over Uhlmann '99 in view of Nyren as applied to claims 1 and 12, and further in view of US Patent 7,078,168 to Sylvan (hereinafter "Sylvan").

Whether claim 35 is obvious under 35 USC §103(a) over Uhlmann '99 in view of Nyren as applied to claim 1 and further in view of US Patent Publication No. 2002/0086324 to Laird (hereinafter "Laird").

Whether claim 37 is obvious under 35 USC §103(a) over Uhlmann'99 in view of Nyren as applied to claim 1 and 8 and further in view of US Patent 5,602,000 to Hyman (hereinafter "Hyman").

ARGUMENT

A. THE COMBINATION OF UHLMANN '99 AND NYREN DOES NOT RENDER CLAIMS 1-5, 7-9, 11-12, 19-20, 22-24, 26-33, 36 AND 37 OBVIOUS

Claims 1-5, 7-9, 11-12, 19-20, 22-24, 26-33 AND 36

Claim 1 is an independent method claim. Claims 2 to 5, 7 to 9, (10), 11, 27, 30, 31, 33 (34), (35), 36, 37 and (39) are directly or indirectly dependent on claim 1.

Claim 12 is an independent method claim. Claims (13-18), 19, 20, 22-24, 26, 28, 29 and (38) are directly or indirectly dependent on claim 12.

Claim 32 is an independent method claim.

The numbers in parenthesis () indicate claims not part of the specified rejection.

The following will show that independent claims 1, 12 ad 32 patentably distinguish the appellant's invention over the combination of Uhlmann and Nyren.

Uhlmann '99 observed differences in 2-D DNA fingerprinting patters in genomic tumor DNA relative to the blood DNA (non-tumor) of certain tumor patients. In one experiment, the authors observed that a "spot" of a particular tumor/non-tumor DNA fragment pair on a 2-D filter was of the same size, but showed different melting behavior. The authors hypothesized that this was the result of different methylation patters at relevant regions of the DNA fragment of interest (Uhlmann '99, page 1748, right col., first full paragraph).

Uhlmann '99 notes that methylation is recognized as an important factor in tumor development as it influences not only the expression of single genes, but also conformation of the DNA and the activity status of a whole chromosome (Uhlmann '99, page 1749, left col., first paragraph).

After seeing fingerprinting patters in methylated and non-methylated test DNAs (phage lamba DNA) similar to that of the DNA of interest, the authors set out to test their

hypothesis, namely whether the differences in the melting characteristics of the spots they had observed were indeed the result of differential methylation.
The authors of Uhlmann '99 decided to determine the methylation status of the DNA fragments of interest using the so-called "bisulfite approach." This technique ("bisulfite –treatment") is based on sodium bisulfite-mediated conversion of non-methylated cytosines to uracil and thus allows the identification of 5-methylcytosine (which is not converted) in genomic (tumor/non-tumor) DNA (Uhlmann '99, page 1749, left col., first full paragraph).

The data presented in Fig. 4 compares corresponding blood and tumor sample pairs and supports that the differences observed in the 2-D fingerprint spots correlate with differential methylation (Uhlmann '99, Fig. 4, page 1752, right column).

Thus, Uhlmann '99 sought out and unveiled that the differences observed in 2-D-fingerprints of bloods and tumor DNA sample pairs could indeed be correlated to differential methylation status of these samples. The authors conclude that 2-D fingerprinting can thus be used to distinguish between methylated and non-methylated DNA of the DNA in question (a DNA fragment which is, in tumor DNA demethylated in the melting domain resulting in the differential 2-D fingerprint patterns: Uhlmann '99, Fig. 2, page 1750, right column).

The Examiner concentrated in her rejection on the specific way with which the authors of Uhlmann '99 test their hypothesis that the differences in the 2-fingerprinting spots they observed are indicative of differential methylation of their tumor and blood (non-tumor) sample.

In particular, Uhlmann '99 proceeds as follows:

- Genomic (blood and tumor) DNA are, after cutting the DNA with restriction enzymes, immobilized in agarose beads to fix the DNA in single stranded form (Uhlmann '99, Fig 1., page 1750, left col., first step). As can be inferred by Figure 4 (Uhlmann '99, Fig. 4, page 1752, right column), blood and tumor DNA are analyzed separately and thus are treated in separate beads.

- The respective DNA is subjected to bisulphite treatment (Uhlmann '99, Fig 1., page 1750, left col., second step).
- After termination of the reaction and washing of the beads the treated DNA of interest, still contained in the beads, is amplified by PCR in separate reactions for the sense and antisense strands of the DNA (see Fig. 1 and description on page 1751, left column).
- The respective amplification products (for sense and antisense strand) are gel extracted and cloned using a Topo TA cloning Kit (INVITROGEN, NL) to produce single stranded DNA.
- The cloned single stranded amplification products are then sequenced by the dideoxynucleotide chain-termination method (see para. 2.5, p. 1750, right column to 1751, right col., l. 3 as well as Fig.1).

Uhlmann '99 is cited for teaching "a method for identifying methylated cytosines comprising treating a sample containing genomic DNA derived from blood and tumor tissue with sodium bisulfite and amplifying the sample by PCR" (Office Action 1/23/09, page 4, lines 4 to 7). The Examiner acknowledged that Uhlmann '99 teaches that the amplified nucleic acids were then cloned and plasmid DNA of the clones was prepared and sequenced using the dideoxynucleotide chain termination method to determine the methylation state of the amplified product.

The Examiner further acknowledged:

- that Uhlmann '99 does not teach a method wherein the amplification primer has a label that forms an anchor for removal of single stranded amplified nucleic acid molecules;
- that Uhlmann '99 does not teach a method wherein said amplification primer is labeled with a biotin.
- that Uhlmann '99 does not teach that the amplified nucleic acids were sequenced using a real-time sequencing method that comprises hyvbridizing a sequencing primer to a single stranded nucleic acid, adding DNA polymerase and other components and detecting a luminescence signal.
- Finally, the Examiner also acknowledged that Uhlmann '99 does not teach a sequencing method that is a high throughput method (Office Action 1/23/09, page 4, last three lines to page 5, line 6)

However, the Examiner argued that Nyren teaches "an alternative method for sequencing. In the method of Nyren, PCR is performed using one or more primers that carry a functional group such as biotin which permits subsequent immobization and aids in the separation of a single stranded DNA (col. 8, lines 1 to 5). Thus, Nyren is said to teach a method wherein the amplification primer had a label that forms an anchor for removal of single stranded amplified nucleic acid molecules" (Office Action 1/23/09, page 5, line 7-8). Nyren is further said to teach real time sequencing. Nyren names many examples in which his method would provide benefits for the user (e.g. Nyren, paragraph bridging col. 13 and 14). None of these examples include a pretreatment of the sample DNA, in particular not with an agent that modifies the DNA and certainly not a pretreatment that would allow one to detect methylation changes in the DNA as a result of the fact that the DNA was changed in a way that the methylated/non-methylated nucleotide show up as different bases in subsequent sequencing.
Nonetheless, the Examiner concluded that it would have been obvious to have modified "the method of Uhlmann by using the sequencing method of Nyren which includes performing PCR with at least one amplification primer labeled with biotin and then sequencing the single stranded nucleic acid via pyrosequencing" (OA 1/23/09; page 6, beginning of first full paragraph).
The Examiner cited a number of advantages that Nyren describes with respect to his method which include high throughput sequencing, an automated approach for large scale sequencing, handling multiple samples in parallel. The Examiner expressed the opinion that the claimed method is obvious "because the substitution of PCR, cloning and sequencing steps performed by Uhlmann for the PCR and sequencing steps performed by Nyren would have been yielded predictable results to one of ordinary still in the art." (OA 1/23/09; page 6, line 22 to page 7, line 2).
Appellants will first argue the specifics of the rejections made and then will discuss the combination of the two references in more general terms.

CLAIMED ELEMENTS NOT ACCOUNTED FOR IN THE OBVIOUSNESS ANALYSIS

The Examiner based her rejection on the rationale that the combination of elements of Uhlmann '99 with Nyren according to known methods yields predictable results (MPEP §2143).

To support this rationale, Office personnel must resolve the *Graham* factual inquiries. Office personnel must, among others, articulate the following:

> (1) a finding that the prior art included each element claimed, although not necessarily in a single prior art reference, with the only difference between the claimed invention and the prior art being the lack of actual combination of the elements in a single prior art reference. (*emphasis added*; MPEP §2143 A. 1.)

Claim 32

The invention as claimed in claim 32 requires "determining the amount of said nucleotide pairing with said new nucleotide pairing partners."

The Examiner has not provided any showing for the element "determining the amount of said nucleotide pairing with said new nucleotide pairing partners" as set forth in the claim and thus has not provided a complete analysis in accordance with MPEP §2143 A. 1.

Claims 1 and 12

The invention as claimed in claims 1 and 12 and all claims dependent thereon requires that treatment of the sample, e.g., with bisulfite, takes place in "an aqueous solution."

The specification clarifies that an "aqueous solution" may be water such as distilled water, a buffered solution such as a phosphate buffered solution or buffered solution other than a phosphate buffered solution, to name some important examples (page 6, lines 6 to 9; page 6, lines 20 to 23; [0016] of the application as published).

The bisulfite treatment in Uhlmann '99 takes place in agarose beads (Uhlmann '99, Fig. 1, page 1750, left column), which facilitates the modification procedure (see description for Fig. 1, page 1750, left column). The DNA remains in Uhlmann '99s beads until after the PCR (amplification step) takes place.

The Examiner has not provided any showing for the element "aqueous solution" as set forth in the claims and thus has not provided a complete analysis MPEP §2143 A. 1.

THE SUBSTITUTION OF PCR, CLONING AND SEQUENCING STEPS PERFORMED BY UHLMANN '99 FOR THE PCR AND SEQUENCING STEPS PERFORMED BY NYREN WOULD NOT HAVE YIELDED PREDICTABLE RESULTS

The Examiner presently argues for the substitution of the PCR, cloning and sequencing steps performed by Uhlmann '99 by the PCR and sequencing steps performed by Nyren.

The substitution of the PCR, cloning and sequencing steps performed by Uhlmann '99 by the PCR and sequencing steps performed by Nyren, thus leaves Uhlmann '99's DNA that is to be subjected to Nyren's PCR and sequencing in the agarose beads (1.7% low melting agarose- in which Uhlmann '99's bisulfite treatment took place and which has a consistency that allows to fix single stranded DNA Uhlmann '99, page 1750, right column, first full paragraph). This might raise the question as to whether the relative complex sequencing reaction of Nyren that follows his PCR could be performed in such an environment. More importantly, the Examiner did not make clear why predictable results should be expected by employing a PCR using detectably labeled amplification primers for subsequent real time sequencing of a DNA sample that is contained in agarose beads. From the teachings of Uhlmann '99, which gel extracts the DNA after PCR, the person skilled in the art would be under the impression that the PCR (amplification) product would need to be gel extracted for further processing, in particular sequencing (Uhlmann '99, page 1750, left column, first paragraph). The Examiner provided no evidence or argument why, despite the lack of gel extraction, the person of ordinary skill in the art would have recognized that the results of the combination were predictable as the Examiner claims (MPEP §2143 A. 3.). Appellants note that an additional step of a gel extraction would be at odds with the identified advantages (speed etc.), that, according to the Office, would cause the person skilled in the art to combine the references.

Appellants have, however, taken note of the fact that Uhlmann '99 teaches an amplification which is followed by gel extraction of the amplification product, cloning to produce single stranded DNA and sequencing.

Taking into account these additional elements taught by Uhlmann '99 in the combination of the Uhlmann '99 and Nyren, appellants would like to point out that Uhlmann '99's amplification primers are not detectably labeled and that Uhlmann '99's amplification product is, after gel extraction, cloned to produce single stranded DNA. The cloning follows Uhlmann '99's amplification and gel extraction and precedes the sequencing. ("[P]lasmid DNA of positive clones . . . were sequenced by the dideoxynucleotide chain-termination method." (Uhlmann '99, see para. 2.5, sentence bridging page 1751, left col. to page 1751, right col.). Appellants note that the person skilled in the art would be reluctant to make the modification to Uhlmann '99's amplification primers, namely detectably label Uhlmann '99's amplification primers, as it would interfere with Uhlmann '99' s subsequent cloning step. For example, U.S. Patent 6,589,736 to Rothschild et al. discloses in its background section, "PCR products that are biotinylated are not suitable material for cloning." (col. 7, starting on line 23). The same patent states also in col. 34, starting on line 40 that "the presence of biotin on the nascent DNA can interfere with its subsequent utilization <u>in cloning</u> or hybridization analysis."

Thus, applicants submit that a modification that employs the PCR as taught by Uhlmann '99 would render Uhlmann '99 unsatisfactory for its intended purpose (see MPEP §2143.01, V- citing *In re Gordon*, 733 F.2d 900, 221 USPQ 1125 (Fed. Cir. 1984).

Furthermore, in *In re Ratti,* where the court noted that the "suggested combination of references would require a substantial reconstruction and redesign of the elements shown in [the primary reference] as well as a change in the *basic principles* under which the [primary reference] construction was designed to operate." *In re Ratti,* 270 F.2d 810, 813 (CCPA 1959) (Emphasis added).

Applicants respectfully submit that the combination of Uhlmann '99 and Nyren change the *basic principles* under which Uhlmann '99 was designed to operate.

Using the current rationale of the Examiner that advocates a replacement of Uhlmann '99's PCR, cloning and sequencing, with Nyren's PCR and sequencing, leaves Uhlmann's DNA in agarose beads (which was used to fix single stranded DNA) for this PCR and sequencing.

Even though the Examiner did not argue that the the bisulfite treatment could have been performed in the solution that Nyren describes and thus a *prima facie* case was not made, e.g., in the paragraph bridging col. 17 and 18, such a change would constitute a substantial reconstruction that would change the basic principle under which Uhlmann '99 was designed to operate.

Taking up a earlier rationale of the Examiner that advocated a modification of Uhlmann 99's PCR with the PCR taught by Nyren, the inclusion of labeled primers in the PCR, would not only render Uhlmann '99 inoperative for its intended purpose, but also constitute a substantial reconstruction that would change the basic principle under which Uhlmann '99 was designed to operate.

Not unlike the fact pattern in *In re Ratti*, the Office seeks to exchange a rather laborious, but "waterproof" method with high speed method. *In re Ratti,* 270 F.2d 810, 813 (CCPA 1959).

THE ADVANTAGES DESCRIBED BY NYREN AND CONSIDERED RELEVANT IN THE OBVIOUSNESS ANALYSIS WOULD NOT MOTIVATE THE PERSON SKILLED IN THE ART TO COMBINE THE TEACHINGS OF UHLMANN '99 AND NYREN

The Examiner relied in her obviousness analysis on the advantages that Nyren describes for his method. Appellant recognize that an advantage is a strong rationale for combining references. (MPEP 2144, II).

The claimed invention is a method for determining the methylation status of a nucleotide. The Examiner uses (a) a reference that seeks to confirm whether differences in 2-D fingerprinting patterns are indeed the result of differential methylation with (b) a reference that provides a new and fast sequencing method.

The question remains if the person would have combined the references considering the teachings of the references in view of the advantages cited by the Office. Uhlmann '99 tried to obtain verification that the differences in 2-D-fingerprinting spots of tumor and non-tumor DNA are in fact a result of changed methylation, a task that requires primarily precision. The prospect of an automatic approach for large scale, non-electrophoretic sequencing procedures which allow for continuous measurements, handling of multiple sample at the same time (1/23/09 Office Action, page 6) as described by Nyren, would be, if at all, at best be of secondary importance.
Nyren himself notes some issues with his method that could affect precision and thus discourage usage in methods that involves a high degree of precision. In col. 7, starting at line 15, Nyren notes:

> "A potential problem which has previously been observed with PPi-based sequencing methods is that DATP, used in the sequencing (chain extension) reaction, interferes in the subsequent luciferase-based detection reaction by acting as a substrate for the luciferase enzyme. <u>This may be reduced or avoided</u> by using, in place of deoxy- or dideoxy adenosine triphosphate (ATP), a DATP or ddATP analogue which is capable of acting as a substrate for a polymerase but incapable of acting as a substrate for a said PPi-detection enzyme." (*emphasis added)*

Furthermore, Nyren also makes clear that accumulation of reaction by-products may take place. While the problem can be avoided by periodic washing, it also adds reluctance if precision is the primary goal as in Uhlmann '99 (Nyren, col. 8, lines 57 to 60). By-products are clearly a concern in Nyren. The additional components of the reaction mixture and additional reaction products stemming from the bisulfite treatment should equally be of concern.

Appellants also note that an obviousness analysis starts with an analysis of the prior art, not from the claimed invention. The question is, whether it would have been obvious, at the time the invention was made, to combine and/or modify the prior art to arrive at the claimed invention.

In this context, appellants direct the Board's attention to the discussion of non-obviousness in *Ortho-McNeil Pharmaceutical v. Mylan Labs*, 520 F.3d 1358, 86 U.S.P.Q.2d 1196 (Fed. Cir. 2008). In particular, appellants direct the Board's attention

to, Judge Rader notation that "In retrospect, Dr. Maryanoff's pathway to the invention, of course, seems to follow the logical steps to produce these properties, but at the time of invention, the inventor's insights, willingness to confront and overcome obstacles, and yes, even serendipity, cannot be discounted." *Id.* at 1364. Hindsight like reasoning is only improper if it included knowledge gleaned only from appellants disclosure. In this context, appellants note that in Ortho-McNeil the Court specifically stated that the TSM test, flexibly applied (in the unpredictable arts) merely assures that the obviousness test proceeds on the basis of evidence – teachings, suggestions (a tellingly broad term), or motivations (an equally broad term) – that arise before the time of invention as the statute requires. Appellants respectfully submit their belief that, for the reasons provided above, the appropriate showing was not provided.

B. THE COMBINATION OF UHLMANN '99, NYREN AND HERMAN DOES NOT RENDER CLAIMS 12-16, 18 AND 38 OBVIOUS;

C. THE COMBINATION OF UHLMANN '99, NYREN, HERMAN AND FEINBERG DOES NOT RENDER CLAIM 17 OBVIOUS

Claim 12 is an independent method claim. Claims 13-16, 17, 18, (19, 20, 22-26, 28 29) and 38 are directly or indirectly dependent on claim 12.

Claims 12 -16 and 18 were rejected over Uhlmann' 99 and Nyren in view of US Patent 6,258,568 to Herman.

Claim 17 was rejected over Uhlmann' 99 and Nyren in view of US Patent 6,258,568 to Herman and further in view of US Publication 2003/0232351 to Feinberg.

With regard to claim 12, the Office explains on page 4 that part of the limitation (d), namely the "to diagnose a pathological condition" or the predisposition therefore is not an actual step, but an intended use of claim limitation (d):

> "(d) detecting whether said nucleotide is methylated or not methylated at said predetermined position in the sample to diagnose said pathological condition or the predisposition for said pathological condition." (*emphasis added*)

Appellants note that this language is presented in the body of the claim and not in the preamble (MPEP §2111.02). Appellants further note that this language does not constitute optional language in accordance with MPEP §2111.04.

The Board is referred to the argument presented with respect to the independent claims.

D. THE COMBINATION OF UHLMANN '99, NYREN AND SYLVAN DOES NOT RENDER CLAIMS 10, 25, 34 AND 39 OBVIOUS

With respect to claims 10, 25, 34 and 39, the Examiner conceded:

- that Uhlmann '99 when combined with Nyren do not teach a method for further comprising calculating a frequency of methylated nucleotides from the results of said real time sequencing (claims 10 and 25).
- that Uhlmann '99 when combined with Nyren do not teach a method wherein an allele frequency of 5% can be detected or a method wherein an allele frequency of 5% with a stardard deviation of no more than 1% is detected (claims 34 and 39).

The Examiner, however, states that US Patent 7,078,168 to Sylvan (hereinafter "Sylan") teaches a method for determining the allele frequency in a population of nucleic acid molecules. The method is in particular used to determine allele frequencies for single nucleotide polymorphisms (SNPs) or other mutations or genetic variations (e.g. nucleotide insertions, additions or deletions, gene, chromosome or genome duplications (or multiplications) etc. in pooled nucleic acid samples or other samples (including single samples)) which may contain allelic variations. Sylan makes no reference to the methylation status of his population of nucleic acid molecules and certainly not "calculating a frequency of methylated nucleotides" as required by claims 10 and 25.

The method of Sylan relies on determining the frequency of an allele in a given population by pooling the nucleic acid sequences of the said population and performing a "primer-extension" type reaction, using primers designed for particular SNPs/alleles. In a sequencing reaction a pattern of nucleotide incorporation in said primer extension products at the positions that correspond to said polymorphic position of interest can be

obtained and the frequency of said allele from said pattern of nucleotide incorporation can be determined.

With regard to claim 34 which is dependent on claim 10, the Office states that due the use of the term "can be" the claim actually does not require detecting an allele frequency.
In this context, appellants note that certain terms may raise the question as to whether it actually limits a claim (MPEP §2111.04). Appellant's respectfully submit that this is not such a case.
In particular, the "can be" term in the context provided clearly states an ability that is either present or not. That is, the method either can detect the allele frequency or not. This ability constitutes a limitation of the claim. The language is, in the context provided, not "conditional" (Office Action, 1/23/09, page 10, last line) and thus constitutes a proper limitation.

With regard to claim 39 which is dependent on claim 10 and omits the term "can be", the Office appears to concede that Sylvan does not exemplify a method wherein an allele frequency of 5% was detected with a standard deviation of not more than 1% (appellant direct the attention to Fig. 6 of Sylvan). However, the Examiner explains that from Sylvan's teaching there is an expectation that an allele frequency of 5% with a standard deviation of not more than 1% could be detected, but suggests that, even if the expected results do not end up being equivalent to the [claimed] results, it would have been obvious to modify the method in order to determine the recited allele frequency. The Examiner has not explained the basis for this "expectation" to support a *prima facie* case of obviousness. The key to supporting any rejection under 35 U.S.C. §103 is the clear articulation of the reason(s) why the claimed invention would have been obvious. The Supreme Court in *KSR* noted that the analysis supporting a rejection under 35 U.S.C. §103 should be made explicit. The Court quoting *In re Kahn,* 441 F.3d 977, 988, 78 USPQ2d 1329, 1336 (Fed. Cir. 2006), stated that "'[R]ejections on obviousness cannot be sustained by mere conclusory statements; instead, there must be some articulated reasoning with some rational underpinning to support the legal conclusion of

obviousness." *KSR Intern. Co. v. Teleflex Inc.,* 550 U.S. 398, 481 (2007) (see §MPEP §2141).

In the obviousness reasoning, the Office merely refers to advantages of Sylvan, in particular the fact that this method "determines the exact sequence of a nucleic acid fragment while directly measuring the amount of nucleotide incorporated." The Office also refers to the accuracy, cost effectiveness and speed with which the information can be obtained.

The Office however, did not provide any reasoning why the person skilled in the art, apart from the advantages of Sylvan's method per se, would make the combination to arrive at the claimed invention, namely calculate the frequency of methylated nucleotides, in particular with the accuracy set forth in claims 34 and 39.

E. THE COMBINATION OF UHLMANN '99, NYREN AND LAIRD DOES NOT RENDER CLAIM 35 OBVIOUS;

F. THE COMBINATION OF UHLMANN '99, NYREN AND HYMAN DOES NOT RENDER CLAIM 37 OBVIOUS

Claim 35 is a method claim dependent upon claim 1. Claim 35 was rejected over Uhlmann '99 in view of Nyren as applied to claim 1 and in further view of Laird.

The Board is referred to the argument presented with respect to claim 1.

Claim 37 is a method claim dependent upon claim 8. Claim 8 is a method claim dependent upon claim 1. Claim 37 was rejected over Uhlmann '99 in view of Nyren as applied to claim 1 and in further view of US Patent 5,602,000 to Hyman.

The Board is referred to the argument presented with respect to claim 1.

CONCLUSION

Having set forth the factual and legal basis which supports the patentability of the claims on appeal, it is respectfully submitted that claims 1 to 5, 7 to 20 and 22 to 39 are allowable. Accordingly, it is respectfully urged that the Board reverse the Examiner's rejection thereof.

Respectfully submitted,

By: /Joyce v. Natzmer/
Joyce von Natzmer
Registration No. 48,120
Customer No. 46002
Direct Line: (301) 657-1282

Pequignot + Myers LLC
200 Madison Ave., 1901
New York, NY 10016

January 25, 2010

CLAIMS APPENDIX

1. A method for detecting the methylation status of a nucleotide at a predetermined position in a nucleic acid molecule comprising:

(a) treating a sample comprising said nucleic acid molecule in an aqueous solution with an agent suitable for the conversion of said nucleotide if present in

(i) methylated form; or

(ii) non-methylated form

(b) to pair with a nucleotide normally not pairing with said nucleotide prior to conversion;

(c) amplifying said nucleic acid molecule treated with said agent via at least one amplification primer to produce an amplification product and converting said amplification product into single stranded amplified nucleic acid molecules, wherein said at least one amplification primer is detectably labeled with a detectable label that forms an anchor for removal of said single stranded amplified nucleic acid molecules to generate a single stranded amplified nucleic acid molecule;

(d) real-time sequencing said single stranded amplified nucleic acid molecule; and

(e) detecting whether said nucleotide is methylated or not methylated at said predetermined position in the sample.

2. The method of claim 1 wherein said sample is derived from a tissue, a body fluid or stool.

3. The method of claim 2 wherein said tissue is a tumor tissue, neurodegenerative tissue or a tissue affected with another neurological disorder.

4. The method of claim 1 wherein said nucleic acid molecule is a DNA molecule or an RNA molecule.

5. The method of claim 1 wherein in (b) the nucleic acid molecule is amplified via LCR or PCR.

6. (Canceled)

7. The method of claim 1 wherein said amplification primer is labeled with (a) biotin, (b) avidin, (c) streptavidin or (d) a derivative of (a), (b) or (c) or a magnetic bead.

8. The method of claim 1 wherein said nucleotide of (a)(i) is an adenine, guanine or a cytosine.

9. The method of claim 1 wherein said real-time sequencing comprises:

(a) hybridization of a sequencing primer to said amplified nucleic acid molecule in single-stranded form;

(b) addition of a DNA polymerase, a ATP sulfurylase, a luciferase, an apyrase, adenosine-phosophosulfate (APS) and luciferin;

(c) sequential addition of dATP, dCTP, dTTP and dGTP;

(d) detection of a luminescent signal wherein an intensity of the luminescent signal is correlated with the incorporation of a specific nucleotide at a specific position in the nucleic acid molecule and wherein the intensity of said signal is indicative of the methylation status of said nucleotide at said predetermined position.

10. The method of claim 1, further comprising calculating a frequency of methylated nucleotides from results of said real-time sequencing.

11. The method of claim 1 wherein said agent suitable for the conversion of said nucleotide to pair with nucleotide normally not pairing with said nucleotide is a bisulfite, preferably sodium bisulfite.

12. A method for the diagnosis of a pathological condition or the predisposition for a pathological condition comprising detection of the methylation status of a nucleotide at a predetermined position in a nucleic acid molecule comprising:

(a) treating a sample comprising said nucleic acid molecule in an aqueous solution with an agent suitable for the conversion of said nucleotide if present in

(i) methylated form; or

(ii) non-methylated form

to pair with a nucleotide normally not pairing with a said nucleotide prior to conversion;

(b) amplifying said nucleic acid molecule treated with said agent via at least one amplification primer to produce an amplification product and converting the amplification product into single stranded amplified nucleic acid molecules, wherein said at least one amplification primer is detectably labeled with a detectable label that forms an anchor for removal of said single stranded amplified nucleic acid molecules to generate a single stranded amplified nucleic acid molecule;

(c) real-time sequencing said single stranded amplified nucleic acid molecule; and

(d) detecting whether said nucleotide is methylated or not methylated at said predetermined position in the sample to diagnose said pathological condition or the predisposition for said pathological condition.

13. The method of claim 12 wherein said pathological condition is cancer, a neurodegenerative disease or another neurological disorder.

14. The method of claim 13 wherein said cancer is a primary tumor, a metastasis or a residual tumor.

15. The method of claim 14 wherein said primary tumor is a glioma.

16. The method of claim 15 wherein said glioma is an astrocytoma, oligodendroglioma, an oligoastrocytoma, a glioblastoma, or a pilocytic astrocytoma.

17. The method of claim 38 wherein said neurodegenerative disease is Alzheimer's disease, Parkinson disease, Huntington disease, or Rett-Syndrome.

18. The method of claim 38 wherein said neurological disorder is Prader-Willi-Syndrome, Angelman-Syndrome, Fragile-X-Syndrome, or ATR-X-Syndrome.

19. The method of claim 12 wherein said nucleic acid molecule is a DNA molecule or an RNA molecule.

20. The method of claim 12 wherein in (b) the nucleic acid molecule is amplified via LCR or PCR.

21. (Canceled)

22. The method of claim 12 wherein said amplification primer is labeled with (a) biotin, (b) avidin, (c) streptavidin or (d) a derivative of (a), (b) or (c) or a magnetic bead.

23. The method of claim 12 wherein said nucleotide of (a)(i) is an adenine, guanine or a cytosine.

24. The method of claim 12 wherein said real-time sequencing comprises:

(a) hybridization of a sequencing primer to said amplified nucleic acid molecule in single-stranded form;

(b) addition of a DNA polymerase, a ATP sulfurylase a luciferase, an apyrase, adenosine-phosphosulfate (APS) and luciferin;

(c) sequential addition of dATP, dCTP, dTTP and dGTP;

(d) detection of a luminescent signal wherein the intensity of the luminescent signal is correlated with the incorporation of a specific nucleotide at a specific position in the nucleic acid molecule and wherein the intensity of said signal is indicative of the methylation status of said nucleotide at said predetermined position.

25. The method of claim 12 further comprising calculating a frequency of methylated nucleotides from results of said real-time sequencing.

26. The method of claim 12 wherein said agent suitable for the conversion of said nucleotide to pair with a nucleotide normally not pairing with said nucleotide is a bisulfite, preferably sodium bisulfite.

27. The method of claim 1 wherein said method is a high-throughput method.

28. The method of claim 12 wherein said sample is derived from tissue, a body fluid or stool.

29. The method of claim 28 wherein said body fluid is blood, serum or urine.

30. The method of claim 1 wherein said nucleotide is a cytosine and is part of one of the following sequences: CpG, CpNpG or CpNpN.

31. The method of claim 1, wherein the methylation status of more than one predetermined nucleotide is detected and a number of samples are analyzed at the same time.

32. A method for generating new nucleotide pairing partners upon amplification of at least one nucleic acid molecule for the detection of the methylation status of nucleotides of said nucleic acid molecule, said method comprising:

(a) providing said at least one nucleic acid molecule;

(b) treating said nucleic acid molecule with an agent suitable for conversion of a nucleotide if present in methylated form or non-methylated form to pair with nucleotide pairing partners normally not pairing with said nucleotide prior to conversion;

(c) amplifying said nucleic acid molecule via at least one amplification primer to produce an amplification product and converting the amplification product into a single stranded nucleic acid molecules, wherein said at

least one amplification primer is detectably labeled with a detectable label that forms an anchor for removal of said single stranded amplified nucleic acid molecules to generate a single stranded amplified nucleic acid molecule comprising said new nucleotide pairing partners normally not pairing with said nucleotide prior to conversion and;

(d) real-time sequencing said single stranded amplified nucleic acid molecule;

(e) determining the amount of said nucleotide pairing with said new nucleotide pairing partners to detect the methylation status of nucleotides of said nucleic acid molecule.

33. The method of claim 1, wherein the methylation status of more than one predetermined nucleotide is determined.

34. The method of claim 10, further comprising detecting an allele frequency, wherein an allele frequency of 5% can be detected.

35. The method of claim 1, wherein said primer does not comprise CpG.

36. The method of claim 1, wherein all nucleotides formerly methylated or not methylated in said nucleic acid molecule are detected.

37. The method of claim 8, wherein said nucleotide of (a)(i) is an adenine or guanine.

38. The method of claim 12, wherein said pathological condition is a neurodegenerative disease or another neurological disorder.

39. The method of claim 10 further comprising detecting an allele frequency of 5% with a standard deviation of not more than 1%.

EVIDENCE APPENDIX

Uhlmann, K. et al., "Changes in methylation patterns identified by two-dimensional DNA fingerprinting," Electrophoresis 20(8):1748-55 (1999): This evidence was entered in the record per "Notice of References Cited" appended to the Office Action of July 17, 2007.

US 6,258,568 (2001) to Nyren et al.: This evidence was entered in the record per "Notice of References Cited" appended to the Office Action of February 22, 2006.

US 5,786,146 (1998) to Herman: This evidence was entered in the record per the "Notice of References Cited" appended to the Office Action of February 22, 2006.

US 2003/0232351 (2003) to Feinberg: This evidence was entered in the record per "Notice of References Cited" appended to the Office Action of February 22, 2006.

US 7,078,168 (2006) to Sylvan This evidence was entered in the record per "Notice of References Cited" appended to the Office Action of July 17, 2007.

US 2002/0086324 (2002) to Laird: This evidence was entered in the record per "Notice of References Cited" appended to the Office Action of July 17, 2007.

US 5,602,000 (1997) to Hyman: This evidence was entered in the record per "Notice of References Cited" appended to the Office Action of June 19, 2008.

Electrophoresis 1999, *20*, 1748–1755

Karen Uhlmann
Karola Marczinek
Jochen Hampe
Gundula Thiel
Peter Nürnberg

Institut für Medizinische Genetik,
Universitätsklinikum Charité,
Berlin, Germany

Changes in methylation patterns identified by two-dimensional DNA fingerprinting

Two-dimensional DNA fingerprinting (2-D fingerprinting) is a sensitive tool for genomic difference analysis between tumor DNA and constitutive DNA of glioma patients. Numerous differences were found even in low-grade gliomas. They can be interpreted as deletions, amplifications, rearrangements, *Hae*III restriction site mutations, tandem repeat instabilities, or methylation differences. The influence of methyl groups on the melting behavior of double-stranded DNA fragments in a denaturing gradient gel was demonstrated by analyzing the migration of λ-phage DNA fragments in 2-D fingerprint gels. A characteristic intensity shift between two neighboring spots in several glioma samples was identified and verified by rehybridization of 2-D filters with a cloned DNA fragment corresponding to the lower spot in 10 out of 11 pilocytic astrocytomas. We hypothesized that this shift may be related to an alteration in the methylation pattern of the tumor DNA. This was specifically tested by analyzing the underlying 750 bp genomic fragment (including 21 CpG dinucleotides) with bisulfite treatment of agarose-embedded DNA. A methylation grade of 88% in tumor DNA as compared to 96% in blood DNA was found. Although only one CpG is located in the melting domain of the cloned fragment, this particular CpG is methylated in all blood samples, but mostly demethylated in the tumor samples. In conclusion, we demonstrate that 2-D fingerprinting may be a powerful tool for the detection of DNA methylation changes in genomic difference analysis.

Keywords: Two-dimensional DNA fingerprinting / Methylation / Bisulfite treatment / Tumor / Pilocytic astrocytoma

EL 3486

1 Introduction

Two-dimensional (2-D) DNA fingerprinting combines two separation techniques. In the first dimension DNA restriction fragments are separated according to their size. In the second dimension separation is achieved by denaturing gradient gel electrophoresis (DGGE), revealing differences in base composition and nucleotide sequence of the fragments. The additional dimension significantly improves the power to detect somatic changes in genomic DNA as compared to one-dimensional DNA fingerprinting [1]. The principle of this methodology was first applied to detect variations in the *Escherichia coli* genome [2]. The introduction of blotting and hybridization with probes specific for repetitive sequences made this method applicable to the analysis of complex eukaryotic genomes [3]. Various applications in different fields of research have been reported [4–10].

Correspondence: Dr. Peter Nürnberg, Institut für Medizinische Genetik, Universitätsklinikum Charité, D-10098 Berlin, Germany
E-mail: peter.nuernberg@charite.de
Fax: +49-30-2802-1286

Abbreviations: ^{5m}**C**, 5-methylcytosine; **dam**, DNA adenosine methylase; **dcm**, DNA cytosine methylase

In a recent study, we used 2-D DNA fingerprinting successfully to screen low-grade gliomas for changes in genomic DNA in comparison to the blood DNA of the patients [11]. Between two and 11 alterations were detected in the 28 blood/tumor pairs analyzed. We believe that these alterations are related to DNA sequences involved in tumor initiation and/or progression. Some of the affected DNA fragments were cloned and rehybridized onto the 2-D filters [12]. For one of these cloned probes, the same spot difference was demonstrated in comparisons of 8 out of 9 independent pilocytic astrocytomas and in 1 of 2 ependymal tumors with the respective constitutive DNAs. A particular spot in the blood DNA pattern that seemed to be split in tumor DNA was observed in these experiments. The two spots were of the same size but showed a different melting behavior, with the additional upper spot in the tumor DNA being immobilized at a lower melting temperature than the original lower spot known from the blood pattern. We interpreted this phenomenon to be a consequence of different methylation patterns of this fragment.

Prokaryotes show methylation at different nucleotides corresponding to a specific pattern of methylases and restriction enzymes. In contrast, in human DNA only cyto-

0173-0835/99/0808-1748 $17.50+.50/0

sines followed by guanine (CpG) appear to be methylated [13]. Cytosine methylation as an epigenetic modification is known to influence a variety of nuclear processes, such as replication of DNA and gene expression [13]. Furthermore, methylation is recognized as an important factor in tumor development [14]. It influences not only the expression of single genes but also the conformation of the DNA strands in extended regions and the activity status of whole chromosomes [15]. Both hypo- and hypermethylation are known to play a crucial role in these regulation processes [13]. As a general rule, a loss of methyl groups leads to increased gene expression [14]. Tumor suppressor genes may be associated with CpG islands, which are not methylated constitutively and thus can be a target of hypermethylation. Oncogenes may also undergo a change in methylation but in these cases a loss of methyl groups is expected to occur in tumorigenesis.

In this study, we present definitive evidence for differential methylation in the blood and tumor of a specific 2-D DNA fingerprinting spot. Initially, 2-D patterns of methylated and nonmethylated λ-DNA were investigated. Usually, the restricted phage genome serves as a standardization marker for the highly variable and complex 2-D patterns of eukaryotic DNA. The methylated λ-DNA showed various "twin spots" whereas in the 2-D pattern of nonmethylated phage DNA the previously characterized 37 single marker spots emerged [16]. This result strengthened our hypothesis and prompted us to determine the methylation state of distinct nucleotides within a human DNA fragment of interest using the bisulfite approach [17]. This technique is based on sodium bisulfite-mediated conversion of nonmethylated cytosines to uracil. It allows the identification of 5-methylcytosines (^{5m}C) in genomic DNA. Only small amounts of material are needed due to subsequent PCR amplification of the modified DNA. After cloning and sequencing of the strand-specific PCR products, this method reveals the methylation status of distinct CpGs in individual DNA strands. We found the tumor DNA of the investigated glioma specific spot to be hypomethylated. Moreover, a distinct CpG within a single-copy sequence showed consistent demethylation in pilocytic astrocytomas and thus may indicate the presence of a candidate oncogene in that region.

2 Materials and methods

2.1 2-D DNA fingerprinting

Genomic DNA from peripheral blood lymphocytes and from tumor tissue was prepared as described previously [11]. Ten micrograms from each DNA sample were digested with 50 units of *Hae*III restriction enzyme according to the supplier's recommendation (Gibco BRL, Eggenstein, Germany). Lambda phage DNA served as marker DNA. The marker was prepared by separate digestions of wild-type methylated λ-DNA from DNA adenosine methylase (*dam*+), DNA cytosine methylase (*dcm*+) *Escherichia coli* Le597 (clind 1 ts857 Sam7; Gibco BRL) and nonmethylated λ-DNA from r_m^-, dam$^-$, and dcm$^-$ *E. coli* GM 119 (Sigma, Deisenhofen, Germany) with the restriction enzymes *Hae*III, *Bgl*I and *Rsa*I according to the supplier's recommendations (Gibco BRL). Next, the λ-DNA fragments were mixed.

The first-dimensional electrophoresis was run in a 6% polyacrylamide gel (PAG; premixed acrylamide/bisacrylamide 37, 5:1, 30% solution; Roth, Karlsruhe, Germany) at 50°C for 4 h at 150 V. The second-dimensional electrophoresis was run at 60°C for 15 h at 150 V in a 6% PAG as well, but with a linear concentration gradient of denaturant included (100% denaturant = 7 M urea, 40% formamide; gradient 10–75%). After gel electrophoresis the DNA was blotted onto a positively charged nylon membrane (Qiabrane; Qiagen, Hilden, Germany) by semidry electroblotting. The obtained filters were hybridized with different mini- and microsatellite core probes. More details about the experimental conditions have been published elsewhere [1, 11].

2.2 Identification of clustered changes

We used the well-defined marker pattern of the restriction enzyme cleaved λ-DNA for standardization. Upon analysis of the spot alterations of 34 patients, a significant clustering of spot changes was observed. The standardization method developed for 2-D DNA fingerprints is described elsewhere [18].

2.3 Elution and cloning of DNA fragments

Spots from 2-D gels were cloned using a protocol which includes the preparation of a duplicate and a master gel [12]. The nylon membrane ("master blot") was used for spot localization in a standard 2-D fingerprinting experiment using the minisatellite probe for the spot cluster of interest. The duplicate diethylaminoethyl (DEAE) membrane was used for spot elution and cloning. The spot position was determined with the help of the λ-DNA marker that was used on both gels. The spot DNA was recovered from the DEAE membrane by high salt elution and amplified by PCR after ligation of adaptor-oligo cassettes. Finally, the PCR products were cloned and used as single locus probes to rehybridize the 2-D filters.

2.4 Mapping of single locus probes

Sequencing of inserts containing the mini- or microsatellite was performed using the Thermo Sequenase cycle

sequencing kit (Amersham Pharmacia Biotech, Cleveland, Ohio, USA). Single-copy sequences flanking the repeat element of the cloned DNA fragment were PCR-amplified and used as probe for a genomic library screening. P1 library filters as well as positive P1 clones were obtained from GenomeSystems (St. Louis, MO, USA). Hybridization of the filters was carried out according to the supplier's instructions. Chromosomal localization of the P1 clones was determined by fluorescence *in situ* hybridization (FISH). Clone DNA was labeled with biotin by nick translation and hybridized to normal human lymphocyte metaphase chromosomes. Regional assignment of signals was determined by analysis of ten well-spread metaphases. A 4′,6-diamidin-2-phenylindol-dihydrochloride (DAPI) counter-staining was performed for band allocation.

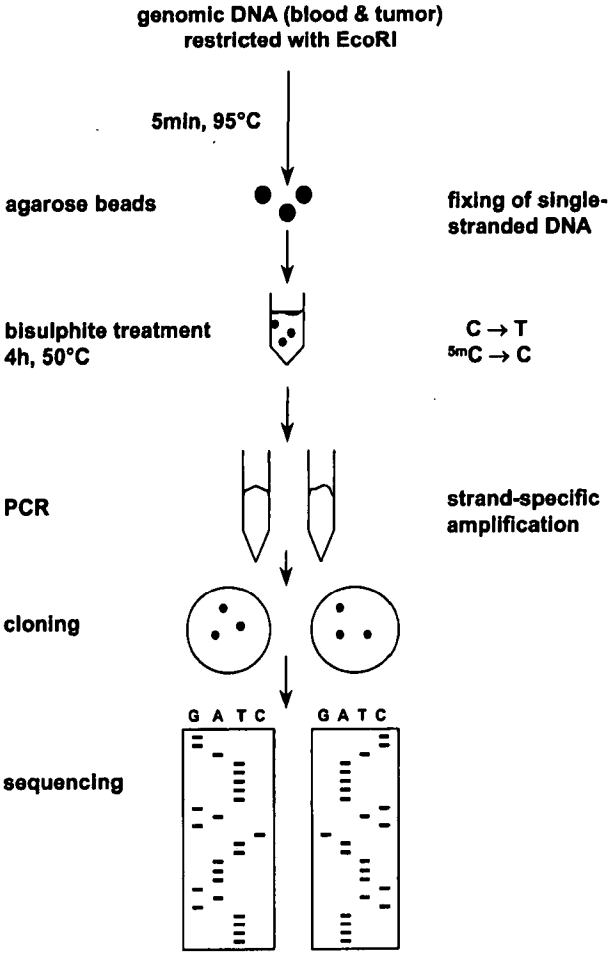


Figure 1. Overview of the experimental approach to determine the methylation state of distinct CpGs. The immobilization of the single-stranded DNA in agarose beads facilitates the modification procedure. After conversion, sense and antisense strands are no longer complementary. Hence, separate PCR reactions are required for the analysis of both strands.

2.5 Determination of ^{5m}C by bisulfite treatment

The protocol is outlined in Fig. 1. Genomic DNA was digested with the restriction enzyme *Eco*RI (Promega, Mannheim, Germany) and precipitated with ethanol. About 100 ng denatured DNA (5 min, 95°C) in 1.7% low melting agarose (Sigma) were dropped into chilled mineral oil to form agarose beads [19]. The fixed single-stranded DNA was subjected to bisulfite treatment (2.5 M sodium metabisulfite, 125 mM hydroquinone, pH 5.0) for 4 h at 50°C. Tubes where covered to protect the reaction against light. After the recommended reaction time, the agarose beads were washed four times in an appropriate

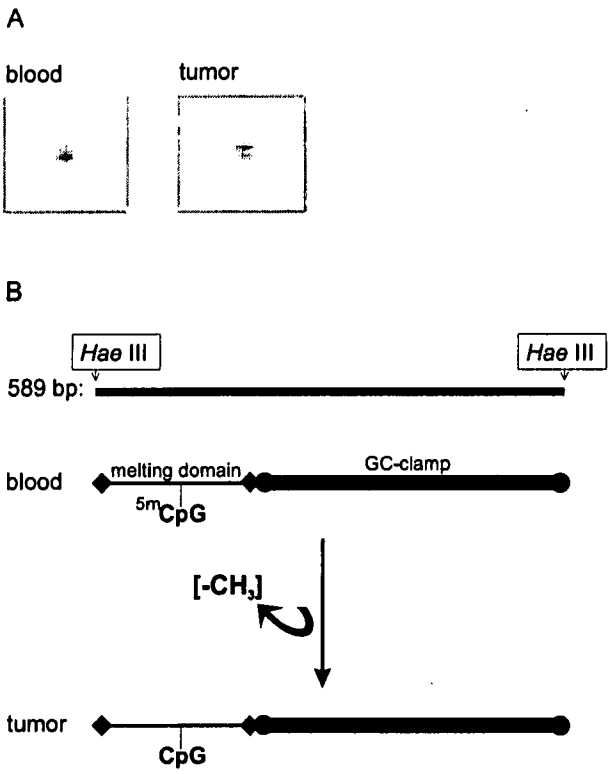


Figure 2. (A) Hybridization of 2-D filters with the single locus probe 48a. The intensity shift obtained with tumor DNA as compared to the patients' blood DNA is typcial for pilocytic astrocytomas and was observed in a number of different patients. Here, the corresponding hybridization signals of patient 493 are shown (*cf.* Fig. 4 in [12]). (B) Schematic presentation of the 2-D DNA fragment detected with probe 48a. *Hae*III is used for digestion of the genomic DNA prior to 2-D electrophoresis. Thus, any fragment is flanked by *Hae*III restriction sites. The fragment contains a large GC-rich repetitive sequence acting as a clamp during the melting process. The lowest melting domain is responsible for the mobility of the fragment in the denaturing gradient. Demethylation of CpGs within this domain might facilitate the melting of the fragment in the tumor DNA.

amount of 1 × TE (10 mM Tris-HCl, pH 8.0, 1 mM EDTA) followed by a desulfonation step in 0.2 M NaOH. This reaction was stopped by adding a 1/5 volume of a 1 M hydrochloric acid solution. Again, the agarose beads were rinsed in 1 × TE and then used for PCR. Beads were stored up to four weeks at 4°C. The sequence of interest was amplified by PCR in separate reactions for the sense and the antisense strand. Each reaction mixture included one agarose bead with about 100 ng of bisulfite-treated DNA in a total volume of 50 µL. PCR reactions were carried out for 40–45 cycles with denaturation at 95°C, annealing at 54–58°C and extension at 72°C. The PCR products were gel extracted (Qiaquick; Qiagen) and cloned using the Topo TA Cloning Kit from Invitrogen (Leek, The Netherlands). Plasmid DNA of positive clones was prepared with Qiaprep (Qiagen) and sequenced by the dideoxynucleotide chain-termination method using the Thermo Sequenase cycle sequencing kit from Amersham Pharmacia Biotech.

3 Results and discussion

3.1 A tumor-related 2-D spot shift specific for pilocytic astrocytomas

A single-locus probe (48a) cloned from a complex 2-D DNA fingerprinting pattern and mapped to 11q14 [12] revealed a typical intensity shift between two neighboring spots when rehybridized to 2-D filters with blood and tumor DNA of different patients (Fig. 2A). Ninety-one percent of all pilocytic astrocytomas screened (10/11) and 11% of the screened astrocytomas (2/18) showed this effect. The observed phenomenon was interpreted as a

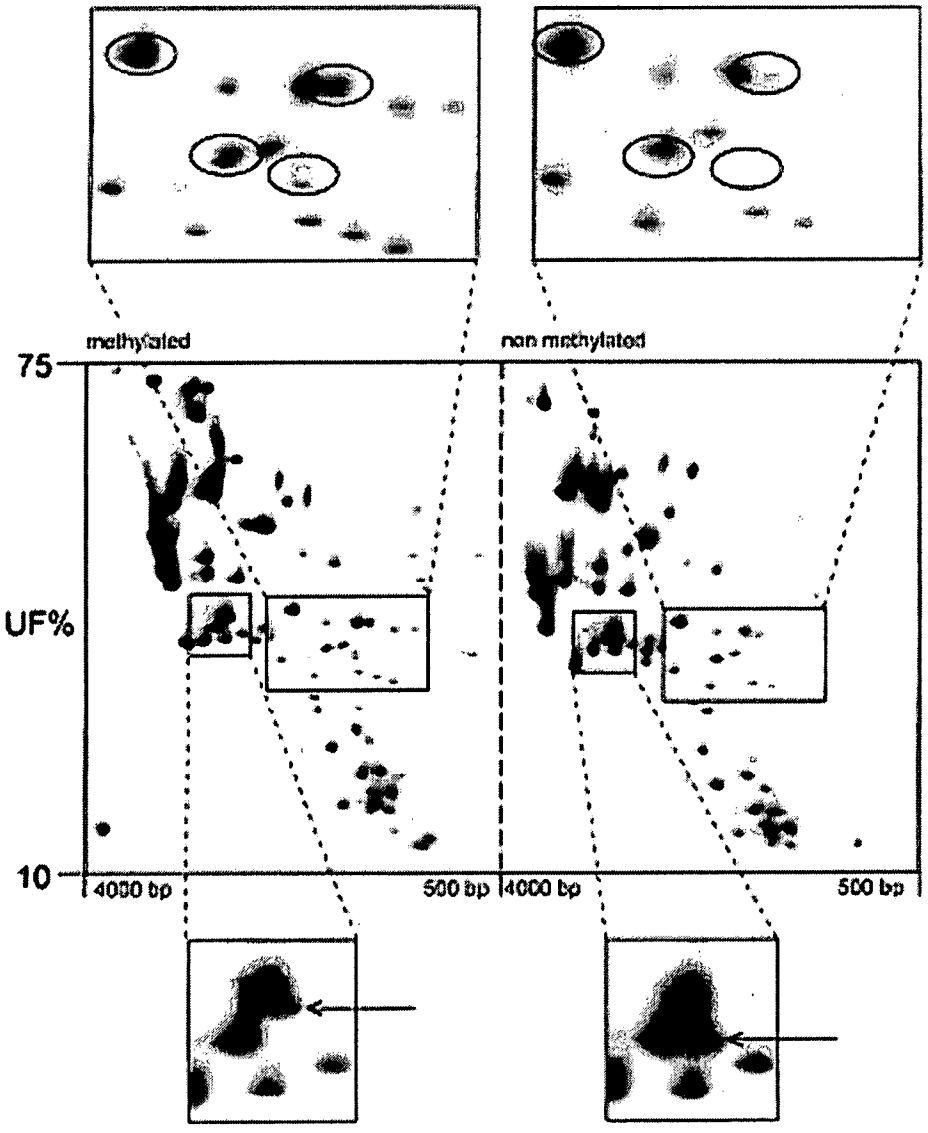


Figure 3. Comparison of methylated and nonmethylated λ-DNA by 2-D DNA fingerprinting. Both DNA samples were analyzed on the same gel to guarantee identical running conditions. Additional upper and lower spots may be due to dam-site and dcm-site methylations, respectively. The upper details show typical twin spots (circled) for the partially methylated DNA. The lower panels reveal a shift of a fragment (arrow) known to possess two adenosine methylation sites in its melting domain. They are thought to be completely methylated in the left sample, thereby lowering the melting temperature of the sequence. %UF, percentage of denaturant (100% denaturant: 7.0 M urea/40% formamide).

consequence of changes in the methylation pattern within the melting domain of the fragment (Fig. 2B). A loss of methyl groups was postulated because the change seemed to lower the melting temperature. This was inferred from experiments performed by Collins and Myers [20] who studied the influence of one or a few methylated bases within the melting domain on the denaturing behavior of this part of the DNA fragment. In human DNA, nearly exclusively cytosine residues in CpGs are methylated [14]. According to Collins and Myers, the loss of this modification in cytosine residues leads to a destabilization of the melting domain and thus to a lower melting temperature (t_m), *i.e.*, a higher spot position within the 2-D gel.

3.2 Influence of methylation on λ-DNA patterns

Results obtained from 2-D experiments performed with methylated and nonmethylated λ-DNA strengthened our hypothesis of a demethylation event in the tumor DNA fragment mentioned above. We separated restriction-digested partially methylated (> 50% of dam and dcm sites) and nonmethylated λ-DNA on a 2-D DNA fingerprinting gel and observed a variety of alterations when comparing the two separation patterns (Fig. 3). In the 2-D DNA fingerprinting pattern of the dam$^+$/dcm$^+$ λ-phage, twin spots and shifts of single spots as compared to the nonmethylated λ-DNA were observed. The double spots indicate sites where the fragments exist in two different methylation states. The single spot shifts may represent fragments turning up only in the modified state if there are active methylation enzymes present. Phage λ possesses 116 dam and 23 dcm sites which have different influences on the melting behavior of the cleaved phage DNA. Methylated adenosines decrease the t_m whereas methylated cytosine residues increase t_m [20]. Upon in-depth analysis of the spots defined in the recent empirical standard pattern analysis [16], we verified the above mentioned influences on spots showing alterations in their separation behavior. All spots of the λ-phage isolated from the dam$^+$/dcm$^+$ *E. coli* strain, which where immobilized earlier in the denaturing gel electrophoresis run than the corresponding spots of the nonmethylated λ-DNA, contained more dam sites than dcm sites (data not shown). However, some of the standard pattern spots did not seem to be immobilized on different denaturant concentrations although they harbored methylation sites as well. A reason for this may be that methylation of certain sites varies with environmental stimuli; thus, these sites may be unmethylated even in the dam$^+$/dmc$^+$ strain [21]. Another explanation is that although the fragments harbor methylation sites, none of these are located in the melting domain and, hence, differential methylation has no impact on the melting behavior.

3.3 Methylation state of the cloned spot DNA fragment

The methylation status of 14 CpGs located in the cloned 589 bp fragment corresponding to probe 48a was investigated by sequencing bisulfite-treated DNA from this fragment (Fig. 4). We chose this method because the PCR amplification steps in the protocol allow for low amounts of DNA to be analyzed. We created two strand-specific primer pairs yielding fragments of 746 and 785 bp, respectively, extending the original cloned sequence by either 100 and 56 bp or 95 and 100 bp at the ends. The flanking sequence information was obtained by sequencing the corresponding P1 clone. The extended sequence contained 21 CpGs, with four of them situated in positions unfavorable for analysis, either too close to an end (CpG No. 1 and 2) or in GC-rich repetitive elements (CpG No. 13 and 14), that are difficult to sequence. Hence, we

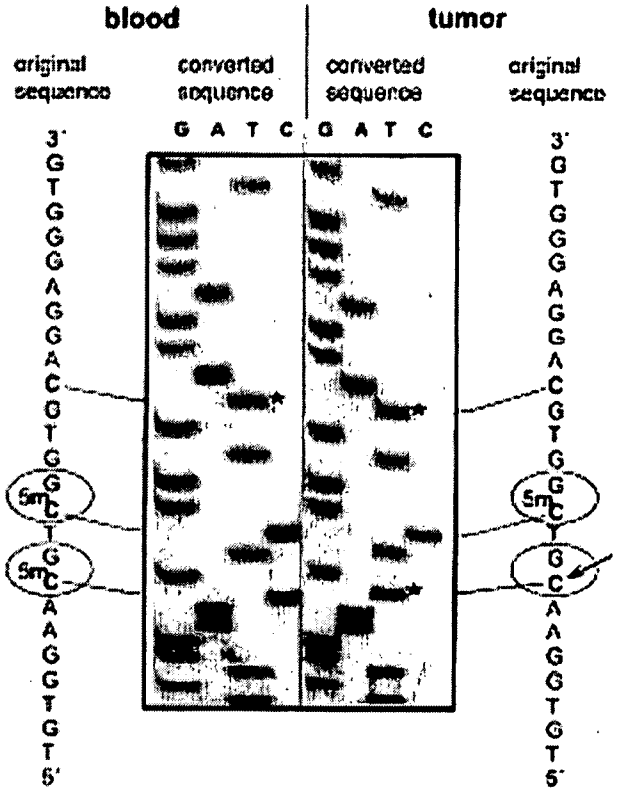


Figure 4. Analysis of differentially methylated CpGs by bisulfite treatment. Sequence ladders of a corresponding blood/tumor sample pair (patient No. 493) after chemical modification are flanked by the original sequences of the untreated DNA. All cytosines appear as thymines (asterisk) unless they were methylated in the original genomic DNA. Two CpGs (circled) are present in the sequence. In the blood DNA both CpGs are methylated. In the tumor DNA, however, the lower one (CpG No. 3) has lost its methyl group (arrow).

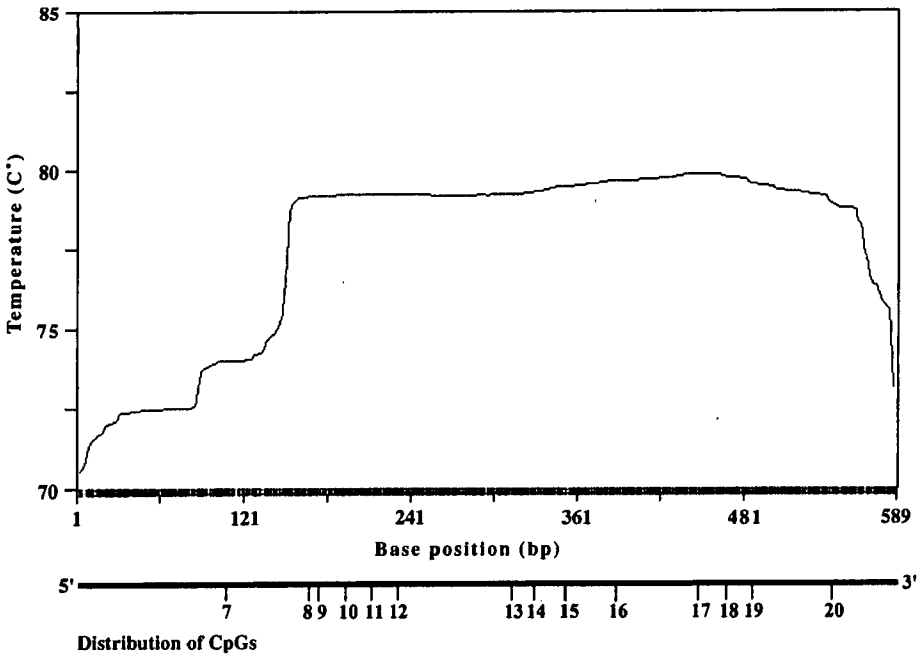


Figure 5. Melting profile of the cloned spot DNA fragment. For each nucleotide position the temperature needed to bring this base pair into an equilibrium of 50% melted and 50% hydrogen bonded molecules is plotted. The three plateaus seen each represent a melting domain. However, the two domains on the left side will in fact behave as one domain because of the minimal difference of only 1–2°C between their melting temperatures. The fragment begins to melt in this region at temperatures at which the long high melting domain still resists melting (GC-clamp). Since the partially melted molecule reduces its electrophoretic mobility drastically, only the sequence of the lower melting domains determines the vertical position of the fragment in the 2-D gel. The distribution of the CpGs given at the bottom reveals that only CpG number 7 is situated in the lower melting domain. Thus, differential methylation of this particular CpG will exclusively influence the initial melting behavior of the fragment.

focused our study on the other 17 CpGs in the sequence. On average, five blood and ten tumor clones of the sense and antisense PCR products were sequenced and analyzed for every patient.

The data for three patients with a pilocytic astrocytoma are summarized in Table 1. Tumor DNA is hypomethylated in comparison to the blood DNA. The average methylation grade in blood was 96% compared to 88% in tumor in the summary over all CpGs in all clones. In patient No. 16026, the same number of different CpGs was found to be demethylated in at least one clone in blood and tumor (*i.e.*, same number of grey cells in Table 1); however, the ratio of methylated to demethylated clones was lower in the tumor. Furthermore, other CpGs were in part demethylated in blood and tumor of that patient. The CpG number 21 was the only one found consistently demethylated in all samples tested, whether blood or tumor. Four CpGs were found to be consistently methylated in blood and tumor, which is to be expected in view of the high level of global methylation in the genomic region. CpG number 7 was consistently methylated in all blood samples but demethylated in all tumor samples. This finding has special significance because this particular CpG is the only one situated in the melting domain of the original fragment cloned from the shifted 2-D DNA fingerprint spot (Fig. 5). Thus, we proved our hypothesis that a tumor-related demethylation event accounts for the observed intensity shift of two neighboring spots in the complex 2-D DNA fingerprint pattern (see Fig. 2). The lower, methylated spot never completely disappeared in the tumor DNA samples. This may be due to contaminating nontumor tissue or incomplete demethylation of CpG number 7 in the tumors. We prefer the former explanation, given our previous experience regarding the purity of tumor samples. In some cases, the blood samples showed a faint upper spot in addition to the prominent lower one. This may indicate that even in normal tissue, methylation of CpG number 7 is not absolutely complete.

A reduced level of global DNA methylation is a common finding in a variety of cancer types [22–24]. Local hypomethylation has been reported to be associated with a higher expression of oncogenes such as c-Ha-ras [25], mage-1 [26], and erb-A1 [27]. A single demethylated CpG can have a significant effect upon the expression of a particular gene, as demonstrated for the CCGG site in the third exon of *c-myc* [28, 29]. Therefore, the use of changes in the methylation pattern as a prognostic marker in cancer is under investigation [30, 31]. We are currently investigating if the methylation status of CpG number 7 (see above) influences the expression of a particular gene. Exon prediction in the surrounding genomic sequence by computer programs (Genie; GenScan) predicted a tentative coding sequence homologous to a family of signal transduction proteins (data not shown).

Table 1. Methylation status of 21 CpGs in a genomic sequence of 750 bp in three patients with pilocytic astrocytomas

CpG	Blood Patient No.			Tumor Patient No.		
No.	454	493	16026	454	493	16026
1	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.
2	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.
3				x	x	
4		x			x	x
5						
6		x	x		x	
7				x	x	x
8				x		
9			x	x	x	
10				x	x	
11	x		x		x	x
12	x				x	
13	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.
14	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.
15		x		x	x	
16				x		
17	x			x		
18						
19						
20						
21	x	x	x	x	x	x

Note: If not indicated otherwise, all clones tested (5 for blood and 10 for tumor) were methylated at this particula CpG.
x, at least one clone was found in which the cytosine was demethylated.

4 Concluding remarks

DNA fingerprinting has been widely used to demonstrate gains and losses of genetic material in tumors as a result of the fatal genetic instability associated with cancer. Here, we have shown that 2-D DNA fingerprinting can also serve as a useful tool in genome-wide screenings for methylation differences between constitutional and tumor DNAs. Furthermore, this is the first report providing indications that hypomethylation may be an important factor for the initiation and/or progression of pilocytic astrocytomas.

We are most grateful to Vera Kalscheuer, Berlin, for helping us to establish the bisulfite treatment method in our laboratory. This work was supported by the Wilhelm Sander Stiftung.

Received March 1, 1999

5 References

[1] Nürnberg, P., Marczinek, K., Thiel, G., Hampe, J., *Electrophoresis* 1995, *16*, 1715–1725.

[2] Fischer, S. G., Lerman, L. S., *Cell* 1979 *16*, 191–200.

[3] Uitterlinden, A. G., Slagboom, P. E., Knook, D. L., Vijg, J., *Proc. Natl. Acad. Sci. USA* 1989, *86*, 2742–2746.

[4] Uitterlinden, A. G., Vijg, J., *Electrophoresis* 1991, *12*, 12–16.

[5] te Meerman, G. J., Mullaart, E., van der Meulen, M. A., den Daas, J. H. G., Morolli, B., Uitterlinden, A. G., Vijg, J., *Am. J. Hum. Genet.* 1993, *53*, 1289–1297.

[6] Hovig, E., Mullaart, E., Borresen, A. L., Uitterlinden, A. G., Vijg, J., *Genomics* 1993, *17*, 67–73.

[7] Verwest, A. M., de Leeuw, W. J. F., Moljin, A. C., Andersen, T. I., Börresen, A. L., Mullaart, E., Uitterlinden, A. G., Vijg, J., *Br. J. Cancer* 1994, *69*, 84–92.

[8] Sidman, C. L., Shaffer, D. J., *Genomics* 1994, *23*, 15–22.

[9] Wu, Y., Hofstra, R. M. W., Scheffer, H., Uitterlinden, A. G., Mullaart, E., Buys, C. H. C. M., Vijg, J., *Hum. Mutat.* 1996, *8*, 160–167.

[10] Wu, Y., Nyström-Lathi, M., Osinga, J., Looman, M. W. G., Peltomäki, P., Aaltonen, L. A., de la Chapelle, A., Hofstra, R. M. W., Buys, C. H. C. M., *Genes Chrom. Cancer* 1997, *18*, 269–278.

[11] Marczinek, K., Hampe, J., Uhlmann, K., Thiel, G., Barth, I., Mrowka, R., Vogel, S., Nürnberg, P., *Glia* 1998, *23*, 130–138.

[12] Marczinek, K., Sugiyama, A., Hampe, J., Thiel, G., Lehmann, K., Neumann, R., de Leeuw, W. J. F., Nürnberg, P., *Electrophoresis* 1997, *18*, 1586–1591.

[13] Baylin, S. B., Herman, J. G., Graff, J. R., Vertino, P. M., Issa, J.-P., *Adv. Cancer Res.* 1998, *72*, 141–196.

[14] Laird, P. W., Jaenisch, R., *Hum. Mol. Genet.* 1994, *3*, 1487–1495.

[15] Strachan, T., Read, A. P., *Human Molecular Genetics*, Bios Scientific Publishers, Oxford 1996.

[16] Hampe, J., Marczinek, K., Preuss, A., Nürnberg, P., *Electrophoresis* 1996, *17*, 659–666.

[17] Frommer, M., McDonald, L. E., Millar, D. S., Collis, C. M., Watt, F., Grigg, G. W., Molloy, P. L., Paul, C. L., *Proc. Natl. Acad. Sci. USA* 1992, *89*, 1827–1831.

[18] Hampe, J., Mrowka, R., Marczinek, K., Nürnberg, P., *Electrophoresis* 1997, *18*, 2874–2879.

[19] Olek, A., Oswald, J., Walter, J., *Nucleic Acids Res.* 1996, *24*, 5064–5066.

[20] Collins, M., Myers, R. M., *J. Mol. Biol.* 1987, *198*, 737–744.

[21] Hale, W. B., van der Woude, M. W., Low, D. A., *J. Bacteriol.* 1994, *176*, 3438–3441.

[22] Gama-Sosa, M. A., Slage, V. A., Trewyn, R. W., Oxenhandler, R., Kuo, K. C., Gehrke, W., Ehrlich, M., *Nucleic Acids Res.* 1983, *11*, 6883–6894.

[23] Goelz, S. E., Vogelstein, B., Hamilton, S. R., Feinberg, A. P., *Science* 1985, *228*, 187–190.

[24] Bernardino, J., Roux, C., Almeida, A., Vogt, N., Gubaud, A., Gerbault-Seureau, M., Magdelenat, H., Bourgeois, C. A., Malfoy, B., Dultrillaux, B., *Cancer Genet. Cytogenet.* 1997, *97*, 83–89.

[25] Feinberg, A. P., Vogelstein, B., *Biochem. Biophys. Res. Commun.* 1983, *111*, 47–54.

[26] De Smet, C., De Backer, O., Faraoni, I., Lurquin, C., Brasseur, F., Boon, T., *Proc. Natl. Acad. Sci. USA* 1996, *93*, 7149–7153.

[27] Lipsanen, V., Leinonen, P., Alhonen, L., Jänne, J., *Blood* 1988, *72*, 2042–2044.

[28] Nambu, S., Inoue, K., Sasaki, H., *Jpn. J. Cancer Res.* 1987, *78*, 695–704.

[29] Sharrard, R. M., Royds, J. A., Rogers, S., Shorthouse, A. J., *Br. J. Cancer* 1992, *65*, 667–672.

[30] Stephenson, J., Akdag, R., Ozbek, N., Mufti, G. J., *Leukemia Res.* 1993, *17*, 291–293.

[31] Belinsky, S. A., Nikula, K. J., Palmisano, W. A., Michels, R., Saccomanno, G., Gabrielson, E., Baylin, S. B., Herman, J. G., *Proc. Natl. Acad. Sci. USA* 1998, *95*, 11891–11896.

US006258568B1

(12) **United States Patent**
Nyren

(10) **Patent No.: US 6,258,568 B1**
(45) **Date of Patent: Jul. 10, 2001**

(54) **METHOD OF SEQUENCING DNA BASED ON THE DETECTION OF THE RELEASE OF PYROPHOSPHATE AND ENZYMATIC NUCLEOTIDE DEGRADATION**

(75) Inventor: **Pal Nyren**, Skarpnack (SE)

(73) Assignee: **Pyrosequencing AB**, Uppsala (SE)

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 0 days.

(21) Appl. No.: **09/331,517**

(22) PCT Filed: **Dec. 22, 1997**

(86) PCT No.: **PCT/GB97/03518**

§ 371 Date: **Jul. 23, 1999**

§ 102(e) Date: **Jul. 23, 1999**

(87) PCT Pub. No.: **WO98/28440**

PCT Pub. Date: **Jul. 2, 1998**

(30) **Foreign Application Priority Data**

Dec. 23, 1996 (GB) 9626815

(51) **Int. Cl.**[7] **C12P 19/34**

(52) **U.S. Cl.** **435/91.1**; 435/91.2

(58) **Field of Search** 435/91.1, 91.2, 435/810

(56) **References Cited**

U.S. PATENT DOCUMENTS

4,851,331	7/1989	Vary et al.	435/6
4,863,849	9/1989	Melamede	435/6
4,971,903	11/1990	Hyman	435/6
5,302,509	4/1994	Cheesemann	435/6
5,405,746	4/1995	Uhlen	435/6
5,498,523	3/1996	Tabor et al.	435/6
5,534,407	7/1996	Tabor et al.	435/5
5,534,424	7/1996	Uhlen et al.	435/91.2

(List continued on next page.)

FOREIGN PATENT DOCUMENTS

3546374	7/1987	(DE) .
414178	6/1993	(DE) .
19602662	8/1997	(DE) .

(List continued on next page.)

OTHER PUBLICATIONS

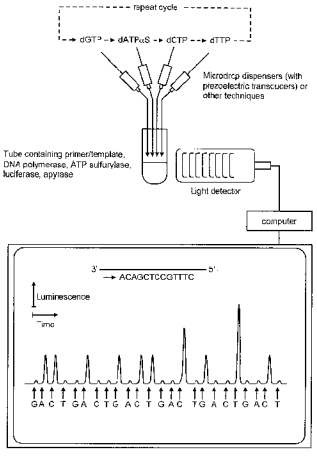
Sanger et al., "DNA sequencing with chain–terminating inhibitors", *Proc. Natl. Acad. Sci USA*, 1977, vol. 74, No. 12, pp. 5463–5467.
STRATAGENE Catalog 1988, 2 pages.
U.S. application No. 09/269436, Nyren et al., filed Jul. 6, 1999.
Fu et al. (1997) Nucleic Acids Research 25:677.
Jones (1997) Bio Techniques 22:938.
Zimmerman (1990) Nucleic Acids Research 18:1067.
Ronaghi et al., Science 281, 363 & 365 (1998).*
Benkovic et al. (1995) Methods in Enzymology 262:257.
Gupta et al. (1984) Nucleic Acids Research 12:5897.
Hultman et al. (1990) Nucleic Acids Research 18:5107.
Hyman (1988) Analytical Biochemistry 174:423.
Kajiyama et al. (1994) Biosci. Biotech. Biochem. 58:1170.
LeBel et al. (1980) J. Biol. Chem. 256:1227.
Nyren (1987) Analytical Biochemistry 167:235.
Nyren (1993) Analytical Biochemistry 208:171.
Nyren et al. (1985) Analytical Biochemistry 151:504.
Patel et al. (1991) Biochemistry 30:511.
Ronaghi et al. (1996) Analytical Biochemistry 242:84.
Syvanen et al. (1990) Genomics 8:684.
Vosberg et al. (1977) Biochemistry 16:3633.
Wong et al. (1991) Biochemistry 30:526.

Primary Examiner—Kenneth R. Horlick
(74) *Attorney, Agent, or Firm*—Baker Botts

(57) **ABSTRACT**

The present invention relates to a method of sequencing DNA, based on the detection of base incorporation by the release of pyrophosphate (PPi) and simultaneous enzymatic nucleotide degradation.

17 Claims, 6 Drawing Sheets



U.S. PATENT DOCUMENTS

5,599,675	2/1997	Brenner et al.	435/6
5,665,545	9/1997	Malek et al.	435/6
5,674,716	10/1997	Tabor et al.	435/91.1
5,679,524	10/1997	Nikiforov et al.	435/6
5,834,189	11/1998	Stevens et al.	435/6
5,849,487	12/1998	Hase et al.	435/6
5,856,092	1/1999	Dale et al.	435/6
5,888,819	3/1999	Goelet et al.	435/5

FOREIGN PATENT DOCUMENTS

0054676	6/1982	(EP) .
0223618	5/1987	(EP) .
0298669	1/1989	(EP) .
0412883	2/1991	(EP) .
0566140	10/1993	(EP) .
0663447	7/1995	(EP) .
0701625	3/1996	(EP) .
0756637	2/1997	(EP) .
2674254	9/1992	(FR) .
WO 8909283	10/1989	(WO) .
WO 8912063	12/1989	(WO) .
9004649	5/1990	(WO) .
9105065	4/1991	(WO) .
WO 9106678	5/1991	(WO) .
WO 9113075	9/1991	(WO) .
9206219	4/1992	(WO) .
9216654	10/1992	(WO) .
WO 9321340	10/1993	(WO) .
9323415	11/1993	(WO) .
WO 9323562	11/1993	(WO) .
WO 9323563	11/1993	(WO) .
WO 9323564	11/1993	(WO) .
WO 9417198	8/1994	(WO) .
9610640	4/1996	(WO) .
9629424	9/1996	(WO) .
9855653	12/1998	(WO) .
9905315	2/1999	(WO) .
0011222	3/2000	(WO) .

* cited by examiner

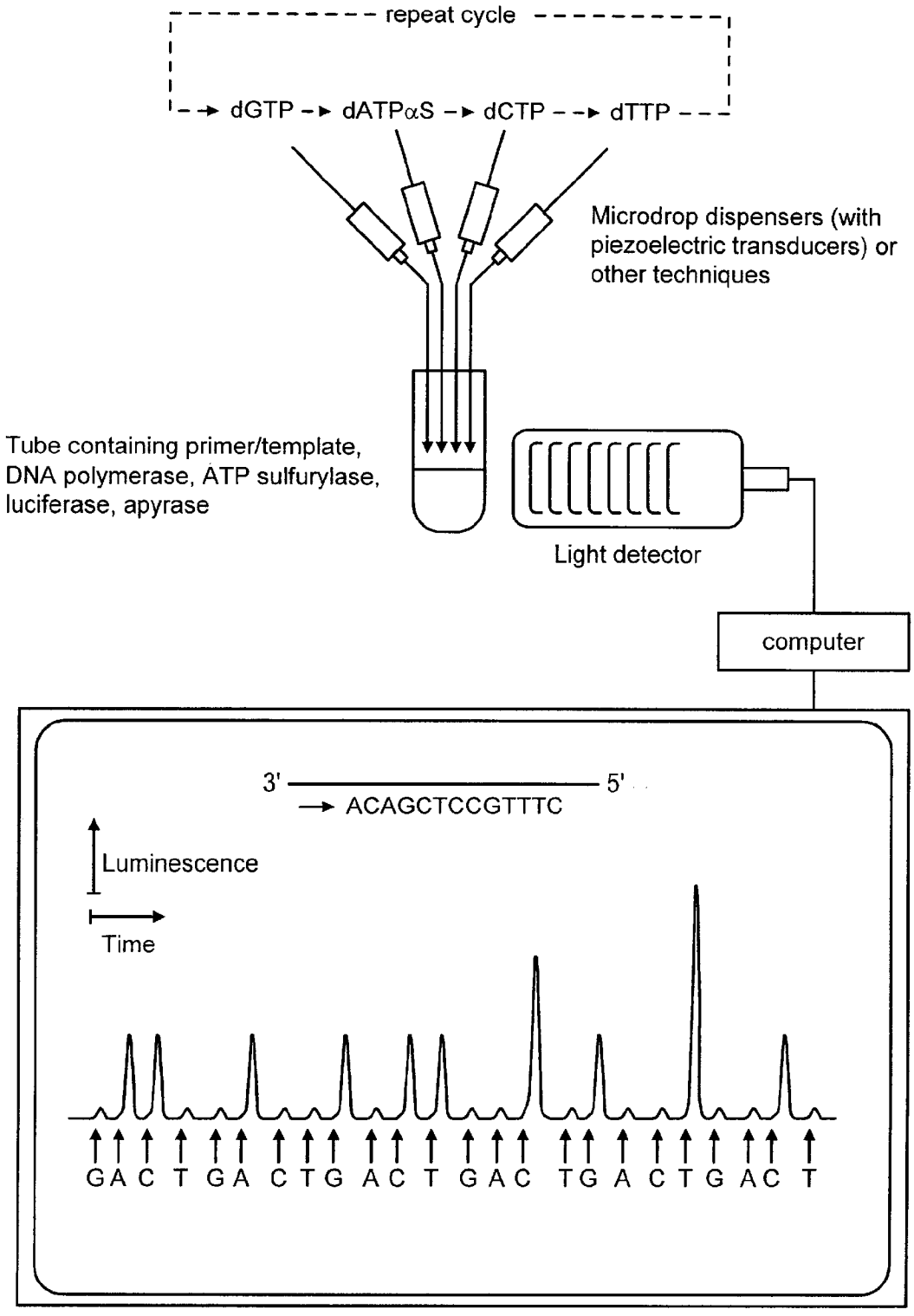

repeat cycle
dGTP
dATPαS
dCTP
dTTP
Microdrop dispensers (with piezoelectric transducers) or other techniques
Tube containing primer/template, DNA polymerase, ATP sulfurylase, luciferase, apyrase
Light detector
computer
3'
5'
ACAGCTCCGTTTC
Luminescence
Time
G A C T G A C T G A C T G A C T G A C T G A C T

FIG. 1

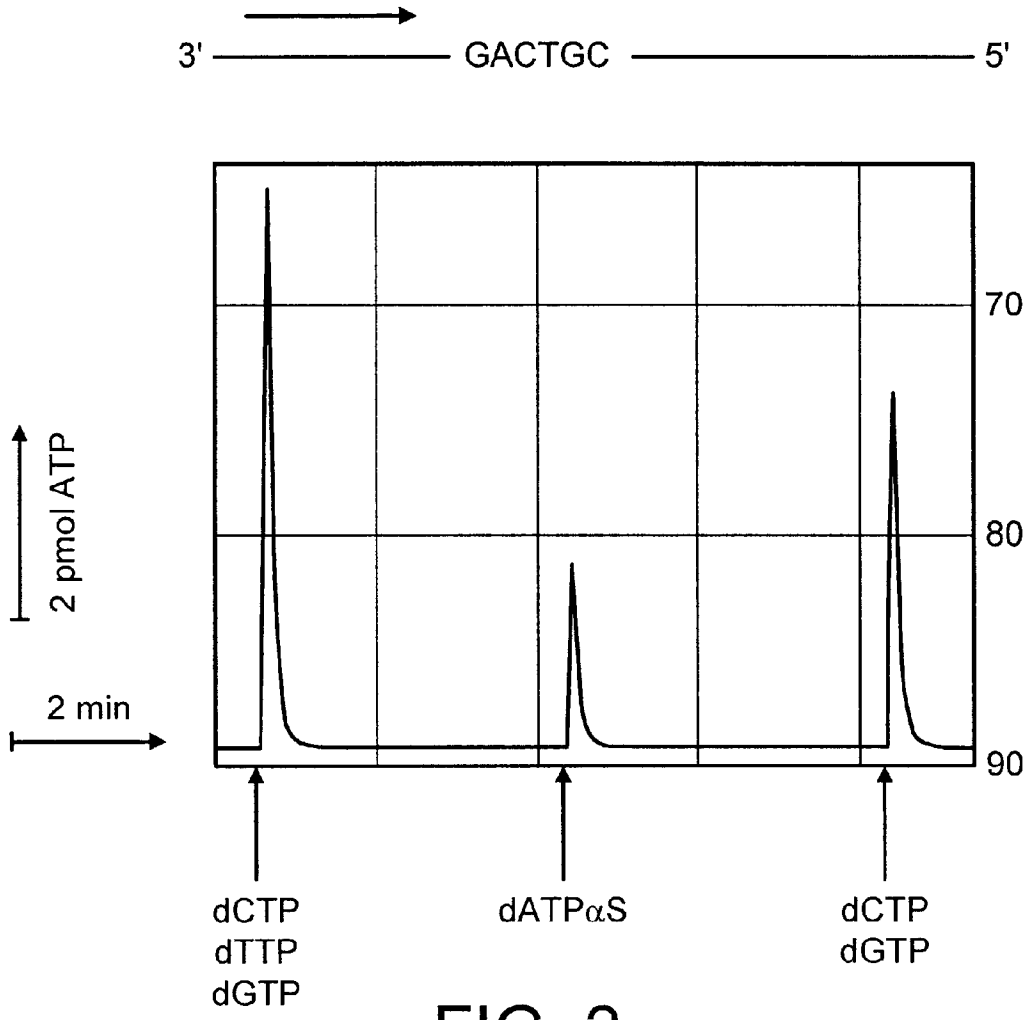

3'
GACTGC
5'
70
80
90
2 pmol ATP
2 min
dCTP
dTTP
dGTP
dATPαS
dCTP
dGTP

FIG. 2

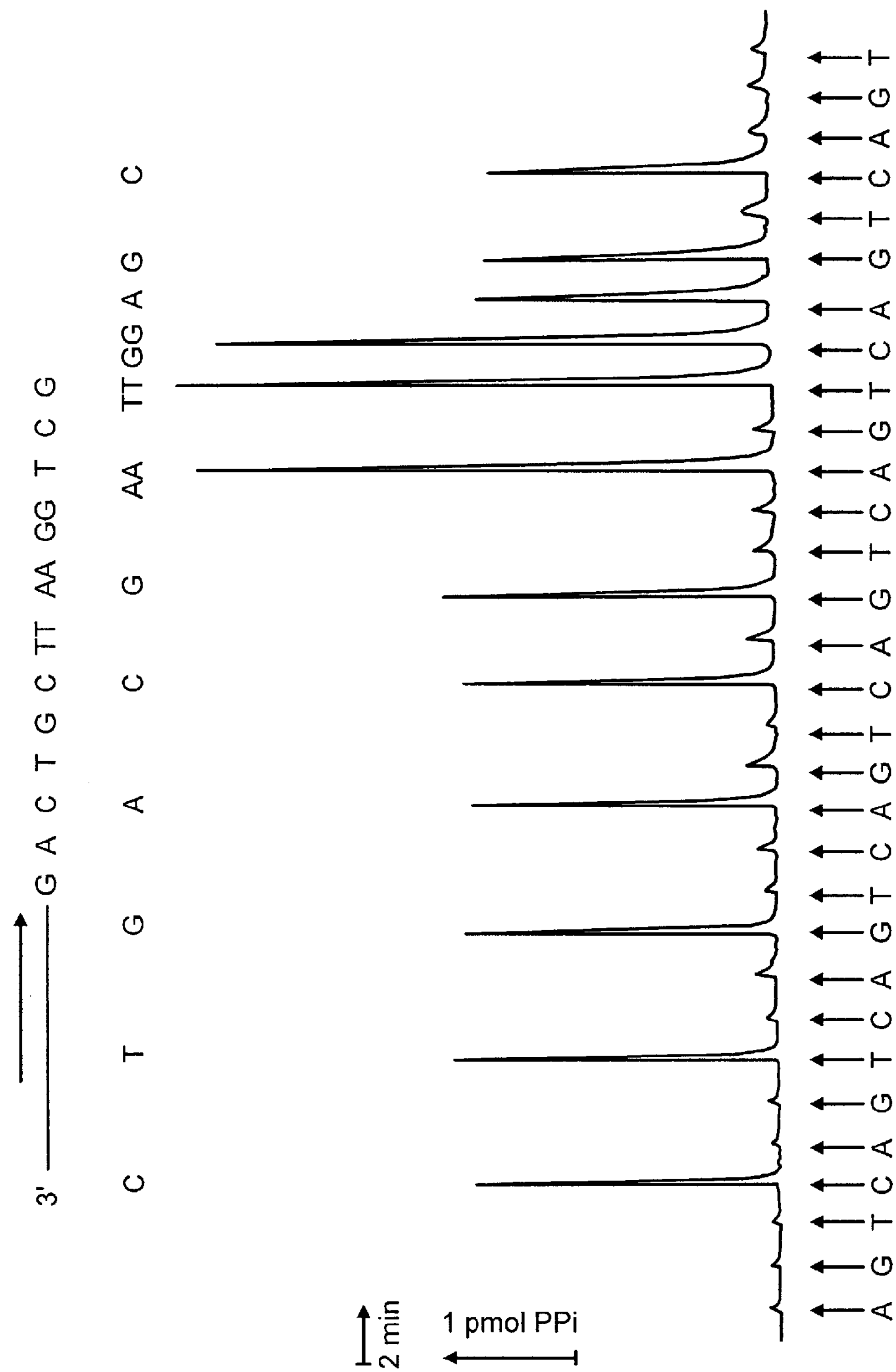

3' GACTGCTTAAGGTCG
C T G A C G AA TT GG A G C
A G T C A G T C A G T C A G T C A G T C A G T C A G T
2 min
1 pmol PPi

FIG. 3

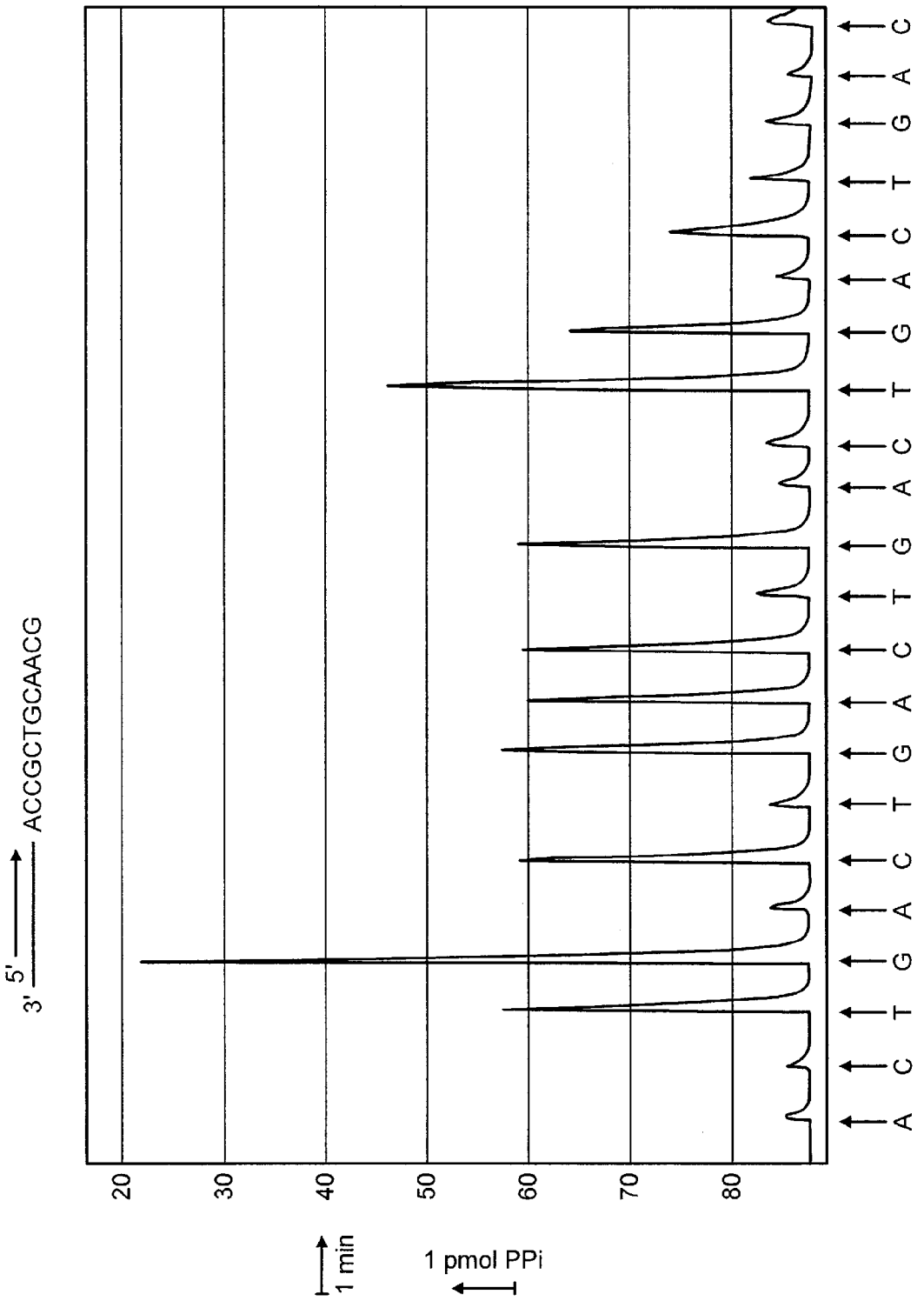

5'
3'
ACCGCTGCAACG
A C T G A C T G A C T G A C T G A C T G A C
20
30
40
50
60
70
80
1 min
1 pmol PPi

FIG. 4

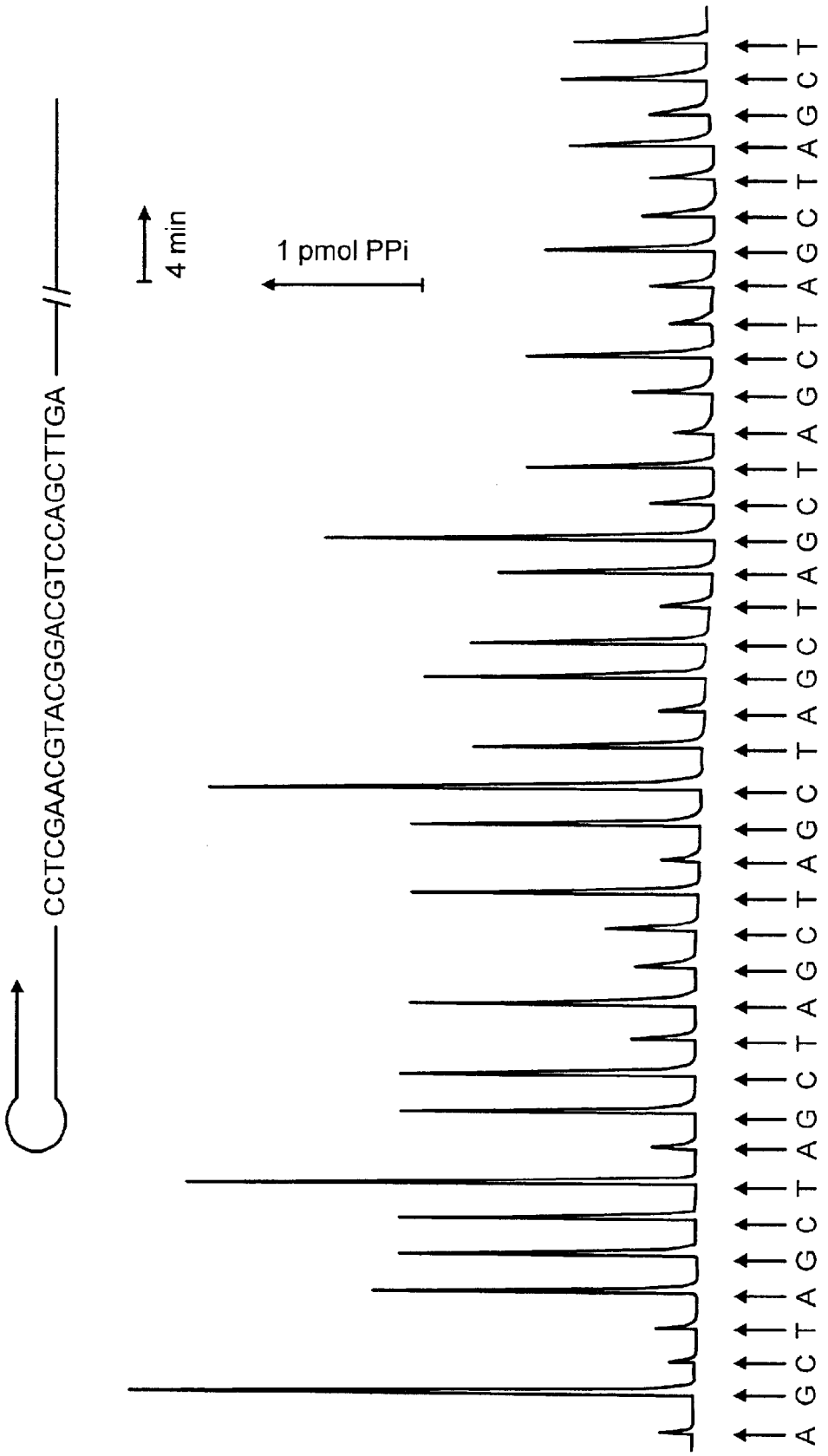

CCTCGAACGTACGGACGTCCAGCTTGA
4 min
1 pmol PPi
A G C T A G C T A G C T A G C T A G C T A G C T A G C T A G C T A G C T

FIG. 5

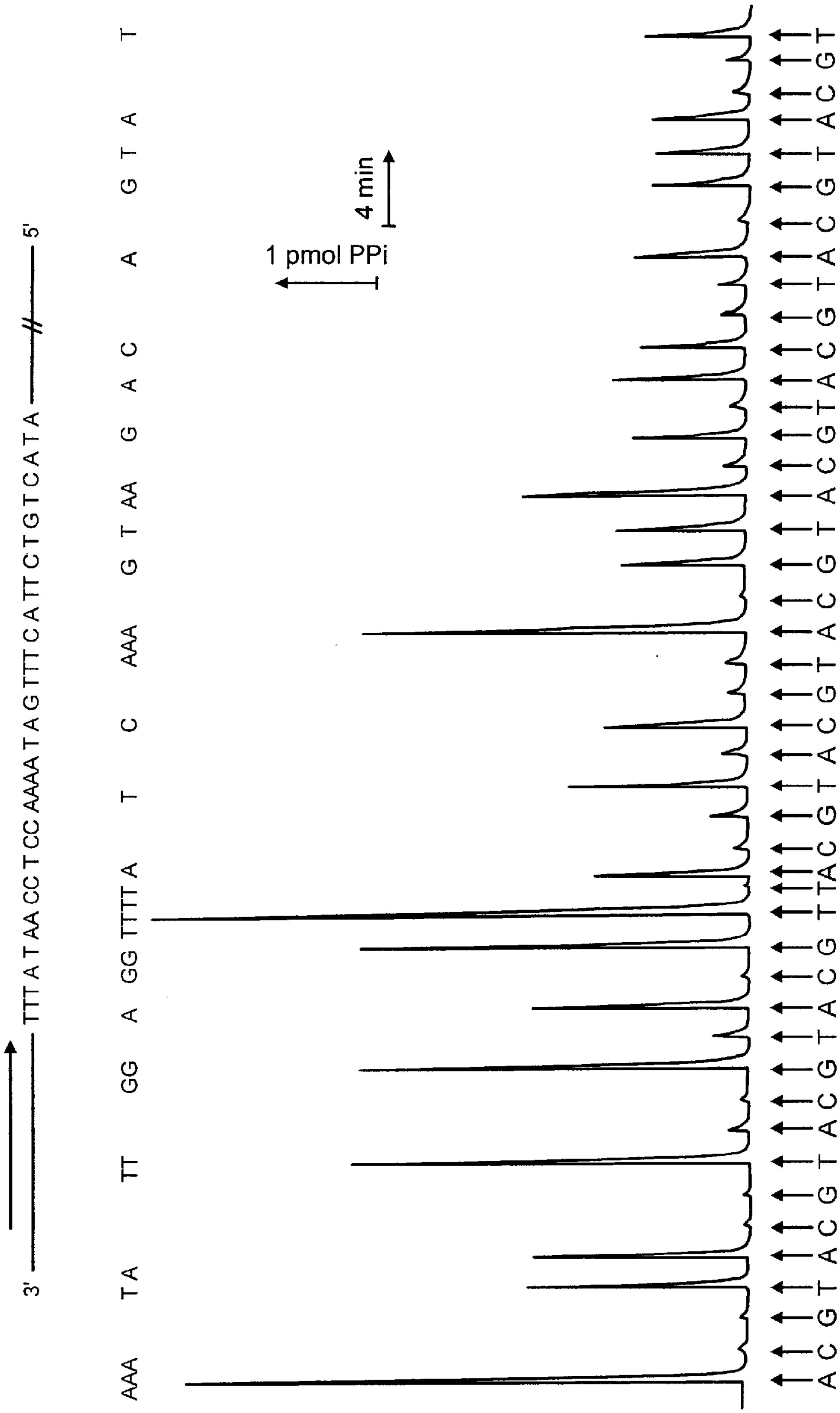

3'
TTTATAACCTCCAAAATAGTTTCATTCTGTCATA
5'
AAA TA TT GG A GG TTTT A T C AAA G T AA G A C A G T A T
1 pmol PPi
4 min
ACGTACGTACGTACGTACGTTACGTACGTACGTACGTACGTACGTACGT

FIG. 6

METHOD OF SEQUENCING DNA BASED ON THE DETECTION OF THE RELEASE OF PYROPHOSPHATE AND ENZYMATIC NUCLEOTIDE DEGRADATION

BACKGROUND OF THE INVENTION

This invention relates to a method of sequencing DNA, based on the detection of base incorporation by the release of pyrophosphate (PPi) and simultaneous enzymatic nucleotide degradation.

DNA sequencing is an essential tool in molecular genetic analysis. The ability to determine DNA nucleotide sequences has become increasingly important as efforts have commenced to determine the sequences of the large genomes of humans and other higher organisms. The two most commonly used methods for DNA sequencing are the enzymatic chain-termination method of Sanger and the chemical cleavage technique of Maxam and Gilbert. Both methods rely on gel electrophoresis to resolve, according to their size, DNA fragments produced from a larger DNA segment. Since the electrophoresis step as well as the subsequent detection of the separated DNA-fragments are cumbersome procedures, a great effort has been made to automate these steps. However, despite the fact that automated electrophoresis units are commercially available, electrophoresis is not well suited for large-scale genome projects or clinical sequencing where relatively cost-effective units with high throughput are needed. Thus, the need for non-electrophoretic methods for sequencing is great and several alternative strategies have been described, such as scanning tunnel electron microscopy (Driscoll et al., 1990, Nature, 346, 294–296), sequencing by hybridization (Bains et al., 1988, J. Theo. Biol. 135, 308–307) and single molecule detection (Jeff et al., 1989, Biomol. Struct. Dynamics, 7, 301–306), to overcome the disadvantages of electrophoresis.

Techniques enabling the rapid detection of a single DNA base change are also important tools for genetic analysis. In many cases detection of a single base or a few bases would be a great help in genetic analysis since several genetic diseases and certain cancers are related to minor mutations. A mini-sequencing protocol based on a solid phase principle was described (Hultman, et al., 1988, Nucl. Acid. Res., 17, 4937–4946; Syvanen et al., 1990, Genomics, 8, 684–692). The incorporation of a radiolabeled nucleotide was measured and used for analysis of the three-allelic polymorphism of the human apolipoprotein E gene. However, radioactive methods are not well suited for routine clinical applications and hence the development of a simple non-radioactive method for rapid DNA sequence analysis has also been of interest.

Methods of sequencing based on the concept of detecting inorganic pyrophosphate (PPi) which is released during a polymerase reaction have been described (WO 93/23564 and WO 89/09283). As each nucleotide is added to a growing nucleic acid strand during a polymerase reaction, a pyrophosphate molecule is released. It has been found that pyrophosphate released under these conditions can be detected enzymically e.g. by the generation of light in the luciferase-luciferin reaction. Such methods enable a base to be identified in a target position and DNA to be sequenced simply and rapidly whilst avoiding the need for electrophoresis and the use of harmful radiolabels.

However, the PPi-based sequencing methods mentioned above are not without drawbacks. The template must be washed thoroughly between each nucleotide addition to remove all non-incorporated deoxynucleotides. This makes it difficult to sequence a template which is not bound to a solid support. In addition new enzymes must be added with each addition of deoxynucleotide.

Thus, whilst PPi-based methods such as are described above do represent an improvement in ease and speed of operation, there is still a need for improved methods of sequencing which allow rapid detection and provision of sequence information and which are simple and quick to perform, lending themselves readily to automation.

We now propose a novel modified PPi-based sequencing method in which these problems are addressed and which permits the sequencing reactions to be performed without intermediate washing steps, enabling the procedure to be carried out simply and rapidly, for example in a single microtitre plate. Advantageously, there is no need to immobilise the DNA. Conveniently, and as will be discussed in more detail below, the new method of the invention may also readily be adapted to permit the sequencing reactions to be continuously monitored in real-time, with a signal being generated and detected, as each nucleotide is incorporated.

BRIEF SUMMARY OF THE INVENTION

In one aspect, the present invention thus provides a method of identifying a base at a target position in a sample DNA sequence wherein an extension primer, which hybridises to the sample DNA immediately adjacent to the target position is provided and the sample DNA and extension primer are subjected to a polymerase reaction in the presence of a deoxynucleotide or dideoxynucleotide whereby the deoxynucleotide or dideoxynucleotide will only become incorporated and release pyrophosphate (PPi) if it is complementary to the base in the target position, any release of PPi being detected enzymically, different deoxynucleotides or dideoxynucleotides being added either to separate aliquots of sample-primer mixture or successively to the same sample-primer mixture and subjected to the polymerase reaction to indicate which deoxynucleotide or dideoxynucleotide is incorporated, characterised in that, a nucleotide-degrading enzyme is included during the polymerase reaction step, such that unincorporated nucleotides are degraded.

BRIEF DESCRIPTION OF THE DRAWINGS

FIG. **1** is a schematic representation of a new DNA sequencing method of the invention. The four different nucleotides are added stepwise to the template hybridised to a primer. The PPi released in the DNA polymerase catalysed reaction, is detected by the ATP sulfurylase and luciferase catalysed reactions. The height of the signal is proportional to the number -of bases which have been incorporated. The added nucleotides are continuously degraded by a nucleotide degrading enzyme. After the first added nucleotide is degraded, the next nucleotide can be added. These steps are repeated in a cycle and the sequence of the template is deduced.

FIG. **2** shows DNA sequencing on a 35-base long oligonucleotide template. About 2 pmol of the template/primer (E3PN/NUSPT) were incubated with 4 pmol (exo$^-$) Klenow and 0.2 U apyrase. The reaction was started by the addition of 0.4 nmol of each of the indicated deoxynucleotides and the PPi released was detected in real-time by the ELIDA. The DNA-sequence of the template is shown in the Figure. The experimental conditions are as described in Example 1.

FIG. **3** shows DNA sequencing on a 35-base-long oligonucleotide template. About 5 pmol of the template/primer (E3PN/NUSPT) were incubated with 8 pmol (exo$^-$) Klenow

and 0.2 U apyrase. The reaction was started by the addition of 0.4 nmol of the indicated deoxynucleotide and the PPi released was detected by the ELIDA. The DNA-sequence of the template is shown in the Figure. The experimental conditions were as described in Example 1.

FIG. **4** shows DNA sequencing on a 35-base-long oligonucleotide template. About 5 pmol of the template/primer (PEBE25/RIT27) were incubated with 8 pmol (exo^-) Klenow and 0.2 U apyrase. The reaction was started by the addition of 0.4 nmol of the indicated deoxynucleotide and the PPi released was detected by the ELIDA. The DNA-sequence of the template is shown in the Figure. The experimental conditions were as described in Example 1.

FIG. **5** shows real-time DNA sequencing performed on a 160-base-long single-stranded PCR product. About 5 pmol of the template/primer (NUSPT) were incubated with 8 pmol (exo^-) Klenow and 0.2 U apyrase. The reaction was started by the addition of 0.4 nmol of the indicated deoxynucleotide and the PPi released was detected by the ELIDA. The DNA-sequence after the primer is shown in the Figure. The experimental conditions were as described in Example 1.

FIG. **6** shows the sequencing method of the invention performed on a 130-base-long single-stranded PCR product hybridized to the sequencing primer as described in Example 2. About 2 pmol of the template/primer was used in the assay. The reaction was started by the addition of 0.6 nmol of the indicated deoxynucleotide and the PPi released was detected by the described method. The DNA-sequence after the primer is indicated in the Figure.

DETAILED DESCRIPTION OF THE INVENTION

The term "nucleotide-degrading enzyme" as used herein includes all enzymes capable of non-specifically degrading nucleotides, including at least nucleoside triphosphates (NTPs), but optionally also di- and mono-phosphates, and any mixture or combination of such enzymes, provided that a nucleoside triphosphatase or other NTP degrading activity is present. Although nucleotide-degrading enzymes having a phosphatase activity may conveniently be used according to the invention, any enzyme having any nucleotide or nucleoside degrading activity may be used, e.g. enzymes which cleave nucleotides at positions other than at the phosphate group, for example at the base or sugar residues. Thus, a nucleoside triphosphate degrading enzyme is essential for the invention. Nucleoside di- and/or mono-phosphate degrading enzymes are optional and may be used in combination with a nucleoside tri-phosphate degrading enzyme. Suitable such enzymes include most notably apyrase which is both a nucleoside diphosphatase and triphosphatase, catalysing the reactions NTP→NMP+2Pi and NTP→NDP+Pi (where NTP is a nucleoside triphosphate, NDP is a nucleoside diphospate, NMP is a nucleotide monophosphate and Pi is phosphate). Apyrase may be obtained from Sigma Chemical Company. Other suitable nucleotide triphosphate degrading enzymes include Pig Pancreas nucleoside triphosphate diphosphohydrolase (Le Bel et al., 1980, J. Biol. Chem., 255, 1227–1233). Further enzymes are described in the literature.

Different combinations of nucleoside tri-, di- or monophosphatases may be used. Such enzymes are described in the literature and different enzymes may have different characteristics for deoxynucleotide degradation, eg. different Km, different efficiencies for a different nucleotides etc. Thus, different combinations of nucleotide degrading enzymes may be used, to increase the efficiency of the nucleotide degradation step in any given system. For example, in some cases, there may be a problem with contamination with kinases which may convert any nucleoside diphosphates remaining to nucleoside triphosphates, when a further nucleoside triphosphate is added. In such a case, it may be advantageous to include a nucleoside disphosphatase to degrade the nucleoside diphosphates. Advantageously all nucleotides may be degraded to nucleosides by the combined action of nucleoside tri-, di- and monophosphatases.

Generally speaking, the nucleotide-degrading enzyme is selected to have kinetic characteristics relative to the polymerase such that nucleotides are first efficiently incorporated by the polymerase, and then any non-incorporated nucleotides are degraded. Thus, for example, if desired the k_m of the nucleotide-degrading enzyme may be higher than that of the polymerase such that nucleotides which are not incorporated by the polymerase are degraded. This allows the sequencing procedure to proceed without washing the template between successive nucleotide additions. A further advantage is that since washing steps are avoided, it is not necessary to add new enzymes eg. polymerase with each new nucleotide addition, thus improving the economy of the procedure. Thus, the nucleotide-degrading enzyme or enzymes are simply included in the polymerase reaction mix, and a sufficient time is allowed between each successive nucleotide addition for degradation of substantially most of the unincorporated nucleotides. The amount of nucleotide-degrading enzyme to be used, and the length of time between nucleotide additions may readily be determined for each particular system, depending on the reactants selected, reaction conditions etc. However, it has for example been found that the enzyme apyrase may conveniently be used in amounts of 0.25 U/mL to 2 U/mL.

As mentioned above, the nucleotide-degrading enzyme(s) may be included during the polymerase reaction step. This may be achieved simply by adding the enzyme(s) to the polymerase reaction mixture prior to, simultaneously with or after the polymerase reaction (ie. the chain extension or nucleotide incorporation) has taken place, e.g. prior to, simultaneously with, or after, the polymerase and/or nucleotides are added to the sample/primer.

In one embodiment, the nucleotide-degrading enzyme(s) may simply be included in solution in a reaction mix for the polymerase reaction, which may be initiated by addition of the polymerase or nucleotide(s).

Alternatively, the nucleotide-degrading enzyme(s) may be immobilised on a solid support e.g. a particulate solid support (e.g. magnetic beads) or a filter, or dipstick etc. and it may be added to the polymerase reaction mixture at a convenient time. For example such immobilised enzyme(s) may be added after nucleotide incorporation (i.e. chain extension) has taken place, and then, when the incorporated nucleotides are hydrolysed, the immobilised enzyme may be removed from the reaction mixture (e.g. it may be withdrawn or captured, e.g. magnetically in the case of magnetic beads), before the next nucleotide is added. The procedure may then be repeated to sequence more bases. Such an arrangement has the advantage that more efficient nucleotide degradation may be achieved as it permits more nucleotide degrading enzyme to be added for a shorter period. This arrangement may also facilitate optimisation of the balance between the two competing reactions of DNA polymerisation and nucleotide degradation.

In a further embodiment, the immobilisation of the nucleotide-degrading enzyme may be combined with the use

of the enzyme(s) in solution. For example, a lower amount may be included in the polymerase reaction mixture and, when necessary, nucleotide-degrading activity may be boosted by adding immobilised enzyme as described above.

The term dideoxynucleotide as used herein includes all 2'-deoxynucleotides in which the 3'-hydroxyl group is absent or modified and thus, while able to be added to the primer in the presence of the polymerase, is unable to enter into a subsequent polymerisation reaction.

PPi can be determined by many different methods and a number of enzymatic methods have been described in the literature (Reeves et al., (1969), Anal. Biochem., 28, 282–287; Guillory et al., (1971), Anal. Biochem., 39, 170–180; Johnson et al., (1968), Anal. Biochem., 15, 273; Cook et al., (1978), Anal. Biochem. 91, 557–565; and Drake et al., (1979), Anal. Biochem. 94, 117–120).

It is preferred to use luciferase and luciferin in combination to identify the release of pyrophosphate since the amount of light generated is substantially proportional to the amount of pyrophosphate released which, in turn, is directly proportional to the amount of base incorporated. The amount of light can readily be estimated by a suitable light sensitive device such as a luminometer.

Luciferin-luciferase reactions to detect the release of PPi are well known in the art. In particular, a method for continuous monitoring of PPi release based on the enzymes ATP sulphurylase and luciferase has been developed by Nyrén and Lundin (Anal. Biochem., 151, 504–509, 1985) and termed ELIDA (Enzymatic Luminometric Inorganic Pyrophosphate Detection Assay). The use of the ELIDA method to detect PPi is preferred according to the present invention. The method may however be modified, for example by the use of a more thermostable luciferase (Kaliyama et al., 1994, Biosci. Biotech. Biochem., 58, 1170–1171) and/or ATP sulfurylase (Onda et al., 1996, Bioscience, Biotechnology and Biochemistry, 60:10, 1740–42). This method is based on the following reactions:

$$PPi + APS \xrightarrow{\text{ATP sulphurylase}} ATP + SO_4^{2-}$$

$$ATP + \text{luciferin} + O_2 \xrightarrow{\text{luciferase}} AMP + PPi + \text{oxyluciferin} + CO_2 + h\nu$$

(APS = adenosine 5'-phosphosulphate)

The preferred detection enzymes involved in the PPi detection reaction are thus ATP sulphurylase and luciferase.

The method of the invention may be performed in two steps, as described for example in WO93/23564 and WO89/09283, firstly a polymerase reaction step ie. a primer extension step, wherein the nucleotide(s) are incorporated, followed by a second detection step, wherein the release of PPi is monitored or detected, to detect whether or not a nucleotide incorporation has taken place. Thus, after the polymerase reaction has taken place, samples from the polymerase reaction mix may be removed and analysed by the ELIDA eg. by adding an aliquot of the sample to a reaction mixture containing the ELIDA enzymes and reactants.

However, as mentioned above, the method of the invention may readily be modified to enable the sequencing (ie. base incorporation) reactions to be continuously monitored in real time. This may simply be achieved by performing the chain extension and detection, or signal-generation, reactions substantially simultaneously by including the "detection enzymes" in the chain extension reaction mixture. This represents a departure from the approach reported in the PPi-based sequencing procedures discussed in the literature above, in which the chain extension reaction is first performed separately as a first reaction step, followed by a separate "detection" reaction, in which the products of the extension reaction are subsequently subjected to the luciferin-luciferase based signal generation ("detection") reactions. This "real time" procedure represents a preferred embodiment of the invention.

To carry out this preferred embodiment of the method of the invention, the PPi-detection enzyme(s) are included in the polymerase reaction step ie. in the chain extension reaction step. Thus the detection enzymes are added to the reaction mix for the polymerase step prior to, simultaneously with or during the polymerase reaction. In the case of an ELIDA detection reaction, the reaction mix for the polymerase reaction may thus include at least nucleotide (deoxy- or dideoxy), polymerase, luciferin, APS, ATP suphurylase and luciferase together with a nucleotide-degrading enzyme. The polymerase reaction may be initiated by addition of the polymerase or, more preferably the nucleotide, and preferably the detection enzymes are already present at the time the reaction is initiated, or they may be added with the reagent that initiates the reaction.

This latter embodiment of the present invention thus permits PPi release to be detected during the polymerase reaction giving a real-time signal. The sequencing reactions may be continuously monitored in real-time. A procedure for rapid detection of PPi release is thus enabled by the present invention. The ELIDA reactions have been estimated to take place in less than 2 seconds (Nyrén and Lundin, supra). The rate limiting step is the conversion of PPi to ATP by ATP sulphurylase, while the luciferase reaction is fast and has been estimated to take less than 0.2 seconds. Incorporation rates for polymerases have also been estimated by various methods and it has been found, for example, that in the case of Klenow polymerase, complete incorporation of one base may take less than 0.5 seconds. Thus, the estimated total time for incorporation of one base and detection by ELIDA is approximately 3 seconds. It will be seen therefore that very fast reaction times are possible, enabling real-time detection. The reaction times could further be decreased by using a more thermostable luciferase. By using a nucleotide-degrading enzyme with a time in the order of seconds for degrading half the nucleotides present, an efficient degradation can be achieved in time frames from seconds to several minutes.

Thus, the method of the present invention may be performed in a single reaction step involving an up to 4-enzyme or more reaction mixture ie. a multi-enzyme mixture. It is surprising that a beneficial and cooperative effect between multiple interlinked enzyme reactions may take place according to the invention and yield beneficial results.

A coupled sequencing/detection system may therefore be based on the following reactions:

$$dNTP + (DNA)_n \xrightarrow{\text{DNA polymerase}} (DNA)_{n+1} + PPi$$

$$PPi \xrightarrow[\text{sulfurylase}]{\text{ATP}} ATP$$

$$ATP \xrightarrow{\text{Luciferase}} \text{Light}$$

$$dNTP \xrightarrow{\text{Apyrase}} dNMP + 2Pi$$

-continued

$$ATP \xrightarrow{\text{Apyrase}} AMP + 2Pi$$

It will be noted that a nucleotide-degrading enzyme such as apyrase would also degrade the ATP not used in the luciferase reactions. Thus, all nucleotide triphosphates are degraded.

Indeed, when PPi release according to the invention is detected by luciferase-based reactions e.g. ELIDA, this ATP-degrading activity may be an important advantage, particularly in "turning off" the light production by the luciferin/luciferase reaction. This may also be of advantage, with a low "burn rate" of the luciferase enzyme.

A potential problem which has previously been observed with PPi-based sequencing methods is that DATP, used in the sequencing (chain extension) reaction, interferes in the subsequent luciferase-based detection reaction by acting as a substrate for the luciferase enzyme. This may be reduced or avoided by using, in place of deoxy- or dideoxy adenosine triphosphate (ATP), a DATP or ddATP analogue which is capable of acting as a substrate for a polymerase but incapable of acting as a substrate for a said PPi-detection enzyme.

The term "incapable of acting" includes also analogues which are poor substrates for the detection enzymes, or which are substantially incapable of acting as substrates, such that there is substantially no, negligible, or no significant interference in the PPi detection reaction.

Thus, a further preferred feature of the invention is the use of a dATP or ddATP analogue which does not interfere in the enzymatic PPi detection reaction but which nonetheless may be normally incorporated into a growing DNA chain by a polymerase and can also be degraded by the nucleotide degrading enzymes. By "normally incorporated" is meant that the nucleotide is incorporated with normal, proper base pairing. In the preferred embodiment of the invention where luciferase is the PPi detection enzyme, the preferred analogues for use according to the invention are the [1-thio] triphosphate (or α-thiotriphosphate) analogues of deoxy or dideoxy ATP, preferably deoxyadenosine [1-thio] triphospate, or deoxyadenosine α-thiotriphosphate (dATPαS) as it is also known. dATPαS, along with the α-thio analogues of dCTP, dGTP and dTTP, may be purchased from New England Nuclear Labs. Experiments have shown that substituting DATP with dATPαS allows efficient incorporation by the polymerase with a low background signal due to the absence of an interaction between dATPαS and luciferase. The signal-to-noise ratio is increased by using a nucleotide analogue in place of dATP, which eliminates the background caused by the ability of dATP to function as a substrate for luciferase. In particular, an efficient incorporation with the polymerase may be achieved while the background signal due to the generation of light by the luciferin-luciferase system resulting from dATP interference is substantially decreased. The dNTPαS analogues of the other nucleotides may also be used in place of all dNTPs.

The sample DNA (ie. DNA template) may conveniently be single-stranded, and may either by immobilised on a solid support or in solution. The use of a nucleotide-degrading enzyme according to the present invention means that it is not necessary to immobilise the template DNA to facilitate washing, since a washing step is no longer required. By using thermostable enzymes, double-stranded DNA templates might also be used.

The sample DNA may be provided by any desired source of DNA, including for example PCR or other amplified fragments, inserts in vectors such as M13 or plasmids.

In order to repeat the method cyclically and thereby sequence the sample DNA and, also to aid separation of a single stranded sample DNA from its complementary strand, the sample DNA may optionally be immobilised or provided with means for attachment to a solid support. Moreover, the amount of sample DNA available may be small and it may therefore be desirable to amplify the sample DNA before carrying out the method according to the invention.

The sample DNA may be amplified, and any method of amplification may be used, for example in vitro by PCR or Self Sustained Sequence Replication (3SR) or in vivo using a vector and, if desired, i vitro and in vivo amplification may be used in combination. Whichever method of amplification is used the procedure may be modified that the amplified DNA becomes immobilised or is provided with means for attachment to a solid support. For example, a PCR primer may be immobilised or be provided with means for attachment to a solid support. Also, a vector may comprise means for attachment to a solid support adjacent the site of insertion of the sample DNA such that the amplified sample DNA and the means for attachment may be excised together.

Immobilisation of the amplified DNA may take place as part of PCR amplification itself, as where one or more primers are attached to a support, or alternatively one or more of the PCR primers may carry a functional group permitting subsequent immobilisation, eg. a biotin or thiol group. Immobilisation by the 5' end of a primer allows the strand of DNA emanating from that primer to be attached to a solid support and have its 3' end remote from the support and available for subsequent hybridisation with the extension primer and chain extension by polymerase.

The solid support may conveniently take the form of microtitre wells, which are advantageously in the conventional 8×12 format, or dipsticks which may be made of polystyrene activated to bind the primer DNA (K Almer, Doctoral Theses, Royal Institute of Technology, Stockholm, Sweden, 1988). However, any solid support may conveniently be used including any of the vast number described in the art, eg. for separation/immobilisation reactions or solid phase assays. Thus, the support may also comprise particles, fibres or capillaries made, for example, of agarose, cellulose, alginate, Teflon or polystyrene. Magnetic particles eg the superparamagnetic beads produced by Dynal AS (Oslo, Norway) also may be used as a support.

The solid support may carry functional groups such as hydroxyl, carboxyl, aldehyde or amino groups, or other moieties such as avidin or streptavidin, for the attachment of primers. These may in general be provided by treating the support to provide a surface coating of a polymer carrying one of such functional groups, e.g. polyurethane together with a polyglycol to provide hydroxyl groups, or a cellulose derivative to provide hydroxyl groups, a polymer or copolymer of acrylic acid or methacrylic acid to provide carboxyl groups or an aminoalkylated polymer to provide amino groups. U.S. Pat. No. 4654267 describes the introduction of many such surface coatings.

Accumulation of reaction by-products may take place. This may readily be avoided by washing the sample after a certain number of reaction cycles e.g. 15–25. Washing may be facilitated by immobilising the sample on a solid surface.

The assay technique is very simple and rapid, thus making it easy to automate by using a robot apparatus where a large number of samples may be rapidly analysed. Since the preferred detection and quantification is based on a luminometric reaction, this can be easily followed spectrophotometrically. The use of luminometers is well known in the art and described in the literature.

The pyrophosphate detection method of the present invention thus opens up the possibility for an automated approach for large-scale, non-elecrophoretic sequencing procedures, which allow for continuous measurement of the progress of the polymerisation reaction with time. The method of the invention also has the advantage that multiple samples may be handled in parallel.

The target DNA may be cDNA synthesised from RNA in the sample and the method of the invention is thus applicable to diagnosis on the basis of characteristic RNA. Such preliminary synthesis can be carried out by a preliminary treatment with a reverse transcriptase, conveniently in the same system of buffers and bases of subsequent PCR steps if used. Since the PCR procedure requires heating to effect strand separation, the reverse transcriptase will be inactivated in the first PCR cycle. When mRNA is the sample nucleic acid, it may be advantageous to submit the initial sample, e.g. a serum sample, to treatment with an immobilised polydT oligonucleotide in order to retrieve all mRNA via the terminal polyA sequences thereof. Alternatively, a specific oligonucleotide sequence may be used to retrieve the RNA via a specific RNA sequence. The oligonucleotide can then serve as a primer for cDNA synthesis, as described in WO 89/0982.

Advantageously, the extension primer is sufficiently large to provide appropriate hybridisation with the sequence immediately 5' of the target position, yet still reasonably short in order to avoid unnecessary chemical synthesis. It will be clear to persons skilled in the art that the size of the extension primer and the stability of hybridisation will be dependent to some degree on the ratio of A-T to C-G base pairings, since more hydrogen bonding is available in a C-G pairing. Also, the skilled person will consider the degree of homology between the extension primer to other parts of the amplified sequence and choose the degree of stringency accordingly. Guidance for such routine experimentation can be found in the literature, for example, Molecular Cloning: a laboratory manual by Sambrook, J., Fritsch E. F. and Maniatis, T. (1989). It may be advantageous to ensure that the sequencing primer hybridises at least one base inside from the 3' end of the template to eliminate blunt-ended DNA polymerase activity. If separate aliquots are used (ie. 4 aliquots, one for each base), the extension primer is preferably added before the sample is divided into four aliquots although it may be added separately to each aliquot. It should be noted that the extension primer may be identical with the PCR primer but preferably it is different, to introduce a further element of specificity into the system.

Alternatively, a primer with a phosphorylated 5'-end, containing a loop and annealing back on itself and the 3'-end of the single stranded template can be used. If the 3'-end of the template has the sequence region denoted T (template), the primer has the following sequence starting from the 5'-end; P-L-P'-T', where P is primer specific (5 to 30 nucleotides), L is loop (preferably 4 to 10 nucleotides), P' is complementary to P (preferably 5 and 30 nucleotides) and T' is complementary to the template sequence in the 3'-end (T) (at least 4 nucleotides). This primer can then be ligated to the single stranded template using T4 DNA ligase or a similar enzyme. This provides a covalent link between the template and the primer, thus avoiding the possibility that the hybridised primer is washed away during the protocol.

The polymerase reaction in the presence of the extension primer and a deoxynucleotide is carried out using a polymerase which will incorporate dideoxynucleotides, e.g. T7 polymerase, Klenow or Sequenase Ver. 2.0 (USB U.S.A.). Any suitable polymerase may conveniently be used and many are known in the art and reported in the literature. However, it is known that many polymerases have a proof-reading or error checking ability and that 3' ends available for chain extension are sometimes digested by one or more nucleotides. If such digestion occurs in the method according to the invention the level of background noise increases. In order to avoid this problem, a nonproof-reading polymerase, eg. exonuclease deficient (exo$^-$) Klenow polymerase may be used. Otherwise it is desirable to add fluoride ions or nucleotide monophosphates which suppress 3' digestion by polymerase. The precise reaction conditions, concentrations of reactants etc. may readily be determined for each system according to choice. However, it may be advantageous to use an excess of polymerase over primer/template to ensure that all free 3' ends are extended.

In the method of the invention there is a need for a DNA polymerase with high efficiency in each extension step due to the rapid increase of background signal which may take place if templates which are not fully extended accumulate. A high fidelity in each step is also desired, which can be achieved by using polymerases with exonuclease activity. However, this has the disadvantage mentioned above that primer degradation can be obtained. Although the exonuclease activity of the Klenow polymerase is low, we have found that the 3' end of the primer was degraded with longer incubations in the absence of nucleotides. An induced-fit binding mechanism in the polymerisation step selects very efficiently for binding of the correct DNTP with a net contribution towards fidelity of 10^5–10^6. Exonuclease-deficient polymerases, such as (exo$^-$) Klenow or Sequenase 2.0, catalysed incorporation of a nucleotide which was only observed when the complementary dNTP was present, confirming a high fidelity of these enzymes even in the absence of proof-reading exonuclease activity. The main advantage of using (exo$^-$) Klenow DNA polymerase over Sequenase 2.0 is its lower Km for nucleotides, allowing a high rate of nucleotide incorporation even at low nucleotide concentrations. It is also possible to replace all dNTPs with nucleotide analogues or non-natural nucleotides such as dNTPαS, and such analogues may be preferable for use with a DNA polymerase having exonuclease activity.

In certain circumstances, e.g. with longer sample templates, it may be advantageous to use a polymerase which has a lower k_m for incorporation of the correct (matched) nucleotide, than for the incorrect (mismatched) nucleotide. This may improve the accuracy and efficiency of the method. Suitable such polymerase enzymes include the α-polymerase of Drosophila.

In many diagnostic applications, for example genetic testing for carriers of inherited disease, the sample will contain heterozygous material, that is half the DNA will have one nucleotide at the target position and the other half will have another nucleotide. Thus if four aliquots are used in an embodiment according to the invention, two will show a negative signal and two will show half the positive signal. It will be seen therefore that it is desirable to quantitatively determine the amount of signal detected in each sample. Also, it will be appreciated that if two or more of the same base are adjacent the 3'-end of the primer a larger signal will be produced. In the case of a homozygous sample it will be clear that there will be three negative and one positive signal when the sample is in four aliquots.

Further to enhance accuracy of the method, bidirectional sequencing ie. sequencing of both strands of a double-stranded template may be performed. This may be advantageous e.g. in the sequencing of heterozygous material. Conveniently, this may be achieved by immobilising the

double-stranded sample template by one strand, e.g. on particles or in a microtitre well, eluting the second strand and subjecting both strands separately to a sequencing reaction by the method of the invention.

In carrying out the method of the invention, any possible contamination of the reagents e.g. the NTP solutions, by PPi is undesirable and may readily be avoided by including a pyrophosphatase, preferably in low amounts, in the reagent solutions. Indeed, it is desirable to avoid contamination of any sort and the use of high purity or carefully purified reagents is preferred, e.g. to avoid contamination by kinases.

Reaction efficiency may be improved by including Mg^{2+} ions in the reagent (NTP and/or polymerase) solutions.

It will be appreciated that when the target base immediately 3'- of the primer has an identical base 3'-thereto, and the polymerisation is effected with a deoxynucleotide (rather than a dideoxynucleotide) the extension reaction will add two bases at the same time and indeed any sequence of successive identical bases in the sample will lead to simultaneous incorporation of corresponding bases into the primer. However, the amount of pyrophosphate liberated will clearly be proportional to the number of incorporated bases so that there is no difficulty in detecting such repetitions.

Since the primer is extended by a single base by the procedure described above (or a sequence of identical bases), the extended primer can serve in exactly the same way in a repeated procedure to determine the next base in the sequence, thus permitting the whole sample to be sequenced.

As mentioned above, in the method of the invention, different deoxy- or dideoxynucleotides may be added to separate aliquots of sample-primer mixture or successively to the same sample-primer mixture. This covers the situations where both individual and multiple target DNA samples are used in a given reaction, which sample DNAs may be the same or different. Thus, for example, as will be discussed in more detail below, in certain embodiments of the invention, there may be one reaction in one container, (in the sense of one sample DNA, ie. one target DNA sequence, being extended) whereas in other embodiments different primer-sample combinations may be present in the same reaction chamber, but kept separate by e.g. area-selective immobilisation.

The present invention provides two principal methods of sequencing immobilised DNA. A. The invention provides a first method of sequencing sample DNA wherein the sample DNA is subjected to amplification; the amplified DNA is optionally immobilised and then subjected to strand separation, one strand eg. the optionally non-immobilised or immobilised strand being removed (ie. either strand may be sequenced), and an extension primer is provided, which primer hybridises to the sample DNA immediately adjacent that portion of the DNA to be sequenced; each of four aliquots of the single stranded DNA is then subjected to a polymerase reaction in the presence of a deoxynucleotide, each aliquot using a different deoxynucleotide whereby only the deoxynucleotide complementary to the base in the target position becomes incorporated; pyrophosphate released by base incorporation being identified. After identification of the incorporated nucleotide a nucleotide degrading enzyme is added. Upon separating the nucleotide degrading enzyme from the different aliquots, for example if it is immobilised on magnetic beads, the four aliquots can be used in a new cycle of nucleotide additions. This procedure can then be continuously repeated. B. The invention also provides a second method of sequencing sample DNA wherein the sample DNA is subjected to amplification; the amplified DNA is optionally immobilised and then subjected to strand separation, one strand eg. the optionally non-immobilised or immobilised strand being removed, and an extension primer is provided, which primer hybridises to the sample DNA immediately adjacent that portion of the DNA to be sequenced; the single stranded DNA is then subjected to a polymerase reaction in the presence of a first deoxynucleotide, and the extent of pyrophosphate release is determined, non-incorporated nucleotides being degraded by the nucleotide-degrading enzyme, and the reaction being repeated by successive addition of a second, third and fourth deoxynucleotide until a positive release of pyrophosphate indicates incorporation of a particular deoxynucleotide into the primer, whereupon the procedure is repeated to extend the primer one base at a time and to determine the base which is immediately 3'- of the extended primer at each stage.

An alternative format for the analysis is to use an array format wherein samples are distributed over a surface, for example a microfabricated chip, and thereby an ordered set of samples may be immobilized in a 2-dimensional format. Many samples can thereby be analysed in parallel. Using the method of the invention, many immobilized templates may be analysed in this way by allowing the solution containing the enzymes and one nucleotide to flow over the surface and then detecting the signal produced for each sample. This procedure can then be repeated. Alternatively, several different oligonucleotides complementary to the template may be distributed over the surface followed by hybridization of the template. Incorporation of deoxynucleotides or dideoxynucleotides may be monitored for each oligonucleotide by the signal produced using the various oligonucleotides as primer. By combining the signals from different areas of the surface, sequence-based analyses may be performed by four cycles of polymerase reactions using the various dideoxynucleotides.

Two-stage PCR (using nested primers), as described in our co-pending application WO90/11369, may be used to enhance the signal to noise ratio and thereby increase the sensitivity of the method according to the invention. By such preliminary amplification, the concentration of target DNA is greatly increased with respect to other DNA which may be present in the sample and a second-stage amplification with at least one primer specific to a different sequence of the target DNA significantly enhances the signal due to the target DNA relative to the 'background noise'.

Regardless of whether one-stage or two stage PCR is performed, the efficiency of the PCR is not critical since the invention relies on the distinct difference different from the aliquots. However, as mentioned above, it is preferred to run an initial qualitative PCR step e.g. by the DIANA method (Detection of Immobilised Amplified Nucleic Acids) as described in WO90/11369 as a check for the presence or absence of amplified DNA.

Any suitable polymerase may be used, although it is preferred to use a thermophilic enzyme such as Taq polymerase to permit the repeated temperature cycling without having to add further polymerase, e.g. Klenow fragment, in each cycle of PCR.

PCR has been discussed above as a preferred method of initially amplifying target DNA although the skilled person will appreciate that other methods may be used instead of in combination with PCR. A recent development in amplification techniques which does not require temperature cycling or use of a thermostable polymerase is Self Sustained

Sequence Replication (3SR). 3SR is modelled on retroviral replication and may be used for amplification (see for example Gingeras, T. R. et al PNAS (USA) 87:1874–1878 and Gingeras, T. R. et al PCR Methods and Applications Vol. 1, pp 25–33).

As indicated above, the method can be applied to identifying the release of pyrophosphate when dideoxynucleotide residues are incorporated into the end of a DNA chain. WO93/23562 relates to a method of identification of the base in a single target position in a DNA sequence (minisequencing) wherein sample DNA is subjected to amplification; the amplified DNA is immobilised and then subjected to strand separation, the non-immobilised strand being removed and an extension primer, which hybridises to the immobilised DNA immediately adjacent to the target position, is provided; each of four aliquots of the immobilised single stranded DNA is then subjected to a polymerase reaction in the presence of a dideoxynucleotide, each aliquot using a different dideoxynucleotide whereby only the dideoxynucleotide complementary to the base in the target position becomes incorporated; the four aliquots are then subjected to extension in the presence of all four deoxynucleotides, whereby in each aliquot the DNA which has not reacted with the dideoxynucleotide is extended to form double stranded DNA while the dideoxy-blocked DNA remains as single stranded DNA; followed by identification of the double stranded and/or single stranded DNA to indicate which dideoxynucleotide was incorporated and hence which base was present in the target position. Clearly, the release of pyrophosphate in the chain terminating dideoxynucleotide reaction will indicate which base was incorporated but the relatively large amount of pyrophosphate released in the subsequent deoxynucleotide primer extension reactions (so-called chase reactions) gives a much larger signal and is thus more sensitive.

It will usually be desirable to run a control with no dideoxynucleotides and a 'zero control' containing a mixture of all four dideoxynucleotides.

WO93/23562 defines the term 'dideoxynucleotide' as including 3'-protected 2'-deoxynucleotides which act in the same way by preventing further chain extension. However, if the 3' protecting group is removable, for example by hydrolysis, then chain extension (by a single base) may be followed by unblocking at the 3' position, leaving the extended chain ready for a further extension reaction. In this way, chain extension can proceed one position at a time without the complication which arises with a sequence of identical bases, as discussed above. Thus, the methods A and B referred to above can be modified whereby the base added at each stage is a 3'-protected 2'-deoxynucleotide and after the base has been added (and the light emission detected), the 3'-blocking group is removed to permit a further 3'-protected - 2' deoxynucleotide to be added. Suitable protecting groups include acyl groups such as alkanol groups e.g. acetyl or indeed any hydroxyl protecting groups known in the art, for example as described in Protective Groups in Organic Chemistry, JFW McOnie, Plenum Press, 1973.

The invention, in the above embodiment, provides a simple and rapid method for detection of single base changes. In one format it successfully combines two techniques: solid-phase technology (DNA bound to magnetic beads) and an Enzymic Luminometric Detection Assay (ELIDA). The method can be used to both identify and quantitate selectively amplified DNA fragments. It can also be used for detection of single base substitutions and for estimation of the heterozygosity index for an amplified polymorphic gene fragment. This means that the method can be used to screen for rare point mutations responsible for both acquired and inherited diseases, identify DNA polymorphisms, and even differentiate between drug-resistant and drug-sensitive strains of viruses or bacteria without the need for centrifugations, filtrations, extractions or electrophoresis. The simplicity of the method renders it suitable for many medical (routine analysis in a wide range of inherited disorders) and commercial applications.

The positive experimental results presented below clearly show the method of the invention is applicable to an on-line automatic non-electrophoretic DNA sequencing approach, with step-wise incorporation of single deoxynucleotides. After amplification to yield single-stranded DNA and annealing of the primer, the template/primer-fragment is used in a repeated cycle of dNTP incubations. Samples are continuously monitored in the ELIDA. As the synthesis of DNA is accompanied by release of inorganic pyrophosphate (PPi) in an amount equal to the amount of nucleotide incorporated, signals in the ELIDA are observed only when complementary bases are incorporated. Due to the ability of the method to determine PPi quantitatively, it is possible to distinguish incorporation of a single base from two or several simultaneous incorporations. Since the DNA template is preferably obtained by PCR, it is relatively straight forward to increase the amount of DNA needed for such an assay.

As mentioned above our results open the possibility for a novel approach for large-scale non-electrophoretic DNA sequencing, which allows for continuous determination of the progress of the polymerisation reaction with time. For the success of such an approach there is a need for high efficiency of the DNA polymerase due to the rapid increase of background signal if templates accumulate which are not "in phase". The new approach has several advantages as compared to standard sequencing methods. Firstly, the method is suitable for handling of multiple samples in parallel. Secondly, relatively cost-effective instruments can be envisioned. In addition, the method avoids the use of electrophoresis and thereby the loading of samples and casting of gels.

A further advantage of the method of the present invention is that it may be used to resolve sequences which cause compressions in the gel-electrophoretic step in standard Sanger sequencing protocols.

The method of the invention may also find applicability in other methods of sequencing. For example, a number of iterative sequencing methods, advantageously permitting sequencing of double-stranded targets, based on ligation of probes or adaptors and subsequent cleavage have been described (see e.g. U.S. Pat. No. 5,599,675 and Jones, BioTechniques 22: 938–946, 1997). Such methods generally involve ligating a double stranded probe (or adaptor) containing a Class IIS nuclease recognition site to a double stranded target (sample) DNA and cleaving the probe/adaptor-target complex at a site within the target DNA, one or more nucleotides from the ligation site, leaving a shortened target DNA. The ligation and cleavage cycle is then repeated. Sequence information is obtained by identifying one or more nucleotides at the terminus of the target DNA. The identification of the terminal nucleotide(s) may be achieved by chain extension using the method of the present invention.

Further to permit sequencing of a double stranded DNA, the method of the invention may be used in a sequencing protocol based on strand displacement, e.g. by the introduc-

tion of nicks, for example as described by Fu et al., in Nucleic Acids Research 1997, 25(3): 677–679. In such a method the sample DNA may be modified by ligating a double-stranded probe or adaptor sequence which serves to introduce a nick e.g. by containing a non- or mono-phosphorylated or dideoxy nucleotide. Use of a strand-displacing polymerase permits a sequencing reaction to take place by extending the 3' end of probe/adaptor at the nick, nucleotide incorporation being detected according to the method of the present invention.

Advantageously, the method according to the present invention may be combined with the method taught in WO93/23563 which uses PCR to introduce loop structures which provide a permanently attached 3' primer at the 3' terminal of a DNA strand of interest. For example, in such a modified method, the extension primer is introduced as part of the 3'-terminal loop structure onto a target sequence of one strand of double stranded DNA which contains the target position, said target sequence having a region A at the 3'-terminus thereof and there being optionally a DNA region B which extends 3' from region A, whereby said double-stranded DNA is subjected to polymerase chain reaction (PCR) amplification using a first primer hybridising to the 3'-terminus of the sequence complementary to the target sequence, which first primer is immobilised or provided with means for attachment to a solid support, and a second primer having a 3'-terminal sequence which hybridises to at least a portion of A and/or B of the target sequence while having at its 5'-end a sequence substantially identical to A, said amplification producing double-stranded target DNA having at the 3'-end of the target sequence, in the following order, the region A, a region capable of forming a loop and a sequence A' complementary to sequence A, whereafter the amplified double-stranded DNA is subjected in immobilised form to strand separation whereby the non-immobilised target strand is liberated and region A' is permitted or caused to hybridise to region A, thereby forming said loop. The 3' end of region A' hybridises immediately adjacent the target position. The dideoxy and/or extension reactions use the hybridised portion as a primer.

The method of the invention may also be used for real-time detection of known single-base changes. This concept relies on the measurement of the difference in primer extension efficiency by a DNA polymerase of a matched over a mismatched 3' terminal. The rate of the DNA polymerase catalyzed primer extension is measured by the ELIDA as described previously. The PPi formed in the polymerization reaction is converted to ATP by ATP sulfurylase and the ATP production is continuously monitored by the firefly luciferase. In the single-base detection assay, single-stranded DNA fragments are used as template. Two detection primers differing with one base at the 3'-end are designed; one precisely complementary to the non-mutated DNA-sequence and the other precisely complementary to the mutated DNA-sequence. The primers are hybridized with the 3'-termini over the base of interest and the primer extension rates are, after incubation with DNA polymerase and deoxynucleotides, measured with the ELIDA. If the detection primer exactly matches to the template a high extension rate will be observed. In contrast, if the 3'-end of the detection primer does not exactly match to the template (mismatch) the primer extension rate will be much lower. The difference in primer extension efficiency by the DNA polymerase of a matched over a mismatched 3'-terminal can then be used for single-base discrimination. Thus, the presence of the mutated DNA sequence can be distinguished over the non-mutated sequence. The relative mismatch extension efficiencies may be strongly decreased by substituting the α-thiotriphosphate analog for the next correct natural deoxynucleotide after the 3'-mismatch termini. By performing the assay in the presence of a nucleotide degrading enzyme. It is easier to distinguish between a match and a mismatch of the type that are easy to extend, such as A:T, T:G and C:T.

The invention also comprises kits for use in methods of the invention which will normally include at least the following components:

(a) a test specific primer which hybridises to sample DNA so that the target position is directly adjacent to the 3' end of the primer;
(b) a polymerase;
(c) detection enzyme means for identifying pyrophosphate release;
(d) a nucleotide-degrading enzyme;
(e) deoxynucleotides, or optionally deoxynucleotide analogues, optionally including, in place of dATP, a dATP analogue which is capable of acting as a substrate for a polymerase but incapable of acting as a substrate for a said PPi-detection enzyme; and
(f) optionally dideoxynucleotides, or optionally dideoxynucleotide analogues, optionally ddATP being replaced by a ddATP analogue which is capable of acting as a substrate for a polymerase but incapable of acting as a substrate for a said PPi-detection enzyme.

If the kit is for use with initial PCR amplification then it will also normally include at least the following components:

(i) a pair of primers for PCR, at least one primer having means permitting immobilisation of said primer;
(ii) a polymerase which is preferably heat stable, for example Taq1 polymerase;
(iii) buffers for the PCR reaction; and
(iv) deoxynucleotides.

The invention will now be described by way of a non-limiting Example with reference to the drawings.

EXAMPLE 1

MATERIALS AND METHODS

Synthesis and purification of oligonucleotides

The oligonucleotides PEBE25 SEQ ID NO:1 (35-mer: 5'-GCAACGTCGCCACACACAACATACGAGCCGGAAGG-3'), RIT 27 SEQ ID NO:2 (23-mer: 5'-GCTTCCGGCTCGTATGTTGTGTG-3'), E3PN SEQ ID NO:3 (35-mer: 5'-GCTGGAATTCGTCAGACTGGCCGTCGTTTTACAAC-3'), NUSPT SEQ ID NO:4 (17-mer: 5'-GTAAAACGACGGCCA-GT-3'), RIT 203 SEQ ID NO:5 (51-mer: 5'-AGCTTGGGTTCGAGGAGATCTTCCGGGTTACGGCGGAAGATCTCCTCGAGG-3), RIT 204 SEQ ID NO:6 (51-mer: 5'-AGCTCC-TCGAGGAGATCTTCCGCCGTAACCCGGAAGATCTCCTCGAACCCA-3'), ROMO 205S SEQ ID NO:7 5'-CGAGGAGATCTTCCGGGTTACGGCG-3), ROMO 205B SEQ ID NO:8 (25-mer: 5'-biotin-CGAGGAGATCTTCCGGGTTACGGCG-3') RIT 28, RIT 29, and USP (Hultman et al., 1990, Nucleic Acids Research, 18, 5107–5112) were synthesised by phosphoramidite chemistry on an automated DNA synthesis apparatus (Gene Assembler Plus, Pharmacia Biotech, Uppsala, Sweden). Purification was performed on a fast protein liquid chromatography pepRPC 5/5 column (Pharmacia Biotech).

In Vitro Amplification and Template Preparation

PCR reactions were performed on the multilinker of plasmid pRT 28 with 7.5 pmol of general primers, RIT 28

and RIT 29 (biotinylated), according to Hultman et al. (supra). The biotinylated PCR products were immobilised onto streptavidin-coated super paramagnetic beads Dynabead™ M280-Streptavidin, or M450-Streptavidin (Dynal, A. S., Oslo, Norway). Single-stranded DNA was obtained by removing the supernatant after incubation of the immobilised PCR product in 0.10 M NaOH for 5 minutes. Washing of the immobilised single-stranded DNA and hybridization to sequencing primers was carried out as described earlier (Nyren et al., 1993, Anal. Biochem. 208, 171–175).

Construction of the Hairpin Vector DRIT 28HP and Preparation of PCR Amplified Template

The oligonucleotides RIT 203, and RIT 204 were hybridized and ligated to HindII (Pharmacia Biotech) pre-restricted plasmid PRIT 28 (the obtained plasmid was named pRIT 28HP). PCR reaction was performed on the multilinker of plasmid pRIT 28HP with 7.5 pmol of primer pairs, RIT 29/ROMO 205S or RIT 27/ROMO 205B, 200 μM DNTP, 20 mM Tris-HCl (pH 8.7), 2 mM $MgCl_2$, 0.1% Tween 20, and 1 unit AmplitTaq DNA polymerase making up a final volume of 50 μl. The temperature profile included a 15 second denaturation step at 95° C. and a 90 second hybridization/extension step at 72° C. These steps were repeated 35 times with a GeneAmp PCR System, 9600 (Perkin, Elmer, Emeryville, USA). The immobilised (as described above) single-stranded DNA obtained from the RIT 29/ROMO 205S amplified reaction or the non-biotinylated single-stranded DNA fragment from the RIT 27/ROMO 205B amplified reaction, was allowed to hybridize at 65° C. for 5 minutes in 20 mM Tris-HCl (pH 7.5), 8 mM $MgCl_2$ to make a self-priming loop structure.

DNA Sequencing

The oligonucleotides E3PN, PEBE25, and the above described PCR products were used as templates for DNA sequencing. The oligonucleotides E3PN, PEBE25, and single-stranded RIT 28/RIT 29 amplified PCR product were hybridized to the primers NUSPT, RIT 27, and NUSPT, respectively. The hybridized DNA-fragments, or the self-primed loop-structures were incubated with either a modified T7 DNA polymerase (Sequenase 2.0; U.S. Biochemical, Cleveland, Ohio, USA), or exonuclease deficient (exo^-) Klenow DNA polymerase (Amersham, UK). The sequencing procedure was carried out by stepwise elongation of the primer strand upon sequential addition of the different deoxynucleoside triphosphates (Pharmacia Biotech), and simultaneous degradation of nucleotides by apyrase. The PPi released due to nucleotide incorporation was detected by the ELIDA. The produced ATP and the non-incorporated deoxynucleotide were degraded in real-time by apyrase. The luminescence was measured using an LKB 1250 luminometer connected to a potentiometric recorder. The luminometer was calibrated to give a response of 10 mV for the internal light standard. The luminescence output was calibrated by the addition of a known amount of ATP or PPi. The standard assay volume was 0.2 ml and contained the following components: 0.1 M Tris-acetate (pH 7.75), 2 mM EDTA, 10 mM magnesium acetate, 0.1% bovine serum albumin, 1 mM dithiothreitol, 2 μM adenosine 5'-phosphosulfate (APS), 0.4 mg/ml polyvinylpyrrolidone (360 000), 100 μg/ml D-100 μg/ml D-luciferin (Bio Therma, Dalarö, Sweden), 4 μg/ml L-luciferin (Bio Therma, Dalarö, Sweden), 120–240 mU/ml ATP sulfurylase (ATP:sulfate adenylyl transferase; EC 2.7.7.4) (Sigma Chemical Co.), 100–400 mU apyrase (nucleoside 5'-triphosphatase and nucleoside 5'-diphosphatase; EC 3.6.1.5) (Sigma Chemical Co.), purified luciferase (Sigma Chemical Co.) in an amount giving a response of 200 mV for 0.1 μM ATP. One to five pmol of the DNA-fragment, and 3 to 15 pmol DNA polymerase were added to the solution described above. The sequencing reaction was started by adding 0.2–1.0 nmol of one of the deoxynucleotides (Pharmacia Biotech). The reaction was carried out at room temperature.

Conventional DNA Sequencing

The sequencing data obtained from the new DNA sequencing were confirmed by semiautomated solid-phase sequencing using radioactive labelled terminators (Hultman et al., 1991, BioTechniques, 10, 84–93). The produced Sanger fragment, from the loop-structured PCR product were restricted by Bgl II restriction endonuclease prior to gel loading.

RESULTS

Principle of the DNA Sequencing Method

The principle of the new sequencing method is illustrated in FIG. **1**. A specific DNA-fragment of interest (sequencing primer hybridized to a single-stranded DNA template, or self-primed single-stranded product) is incubated with DNA polymerase, ATP sulfurylase, luciferase and a nucleotide degrading enzyme, and a repeated cycle of nucleotide incubation is performed. The synthesis of DNA is accompanied by release of PPi equal in molarity to that of the incorporated nucleotide. Thereby, real-time signals are obtained by the enzymatic inorganic pyrophosphate detection assay (ELIDA) only when complementary bases are incorporated. In the ELIDA the produced PPi is converted to ATP by ATP sulfurylase and the amount of ATP is then determined by the luciferase assay (FIG. **1**). As added nucleoticles are continuously degraded by a nucleotide degrading enzyme a new nucleotide can be added after a specific time-interval. From the ELIDA results the sequence after the primer is deduced. The DNA sequencing method of the invention is named "pyrosequencing".

Optimization of the Method

Several different parameters of the new DNA sequencing approach were optimised in a model system using a synthetic DNA template. As the method is based on utilization, of added deoxynucleotides by the DNA polymerase detection of released PPi by a coupled enzymatic system and continuous degradation of nucleotides, the concentration of the different components used in the assay should be carefully balanced.

The signal-extent as a function of the numbers of correct deoxynucleotides added is shown in FIG. **2**. The reaction was started by addition of the three first correct bases (dCTP, dTTP and dGTP) and the trace show both the release of PPi (converted to ATP by the ATP sulfurylase) during the incorporation of the bases, and the subsequent degradation of ATP. The incorporation of three residues was noted. After a short time-lag (the apyrase reaction was allowed to proceed about 2 minutes), dATPuS was added; a signal corresponding to incorporation of one residue was observed. Thereafter, the two next correct deoxynucleotides (dCTP and dGTP) were added. This time the incorporation of two residues was detected. The results illustrated in FIG. **2** show that the DNA sequencing approach functions; the added deoxynucleotides were degraded by apyrase between each addition, the observed signals were proportional to the amount of nucle-

otide incorporated, and no release of PPi was observed if a non-complementary base was added (not shown).

In the above illustrated experiment, 32 mU ATP sulfurylase, 200 mU apyrase, 2 U (exo$^-$) Klenow, 2 pmol template/primer, and 0.4 pmol deoxynucleotides, were used. Similar results were obtained (not shown) when the different compounds were varied within the interval: 24–48 mU ATP sulfurylase, 100–400 mU apyrase, 1–5 U (exo$^-$) Klenow, 1–5 pmol template/primer, and 0.2–1.0 nmol deoxynucleotides. It may be important to use an excess of polymerase over primer/template to be sure that all free 3' ends are extended. It may also be important that the sequencing primer hybridize at least one base inside from the 3' end of the template to eliminate blunt-end DNA polymerase activity (Clark, 1991, Gene, 104, 75–80).

DNA Sequencing

In the next series of experiments two different synthetic templates as well as a PCR product were sequenced in order to investigate the feasibility of the new approach. FIGS. **3** and **4** show the result from DNA sequencing performed on two different synthetic templates. Both templates were sequenced to the end, and in both cases the true sequence could be determined. When the polymerase reaches the end of the template, the signal strongly decreases indicating slower polymerization for the last bases. The signal was not decreased to the same extent if a longer template was sequenced (FIG. **5**). The small signals observed when non-complementary bases were added are due to PPi contamination in the nucleotide solutions. The later increase of this background signal (false signals) is probably due to nucleoside diphosphate kinase activity (contamination in the ATP sulfurylase preparation from Sigma). The nucleoside diphosphate kinase converts non-degraded deoxynucleoside diphosphates to deoxynucleoside triphosphates when a new deoxynucleotide triphosphate is added. The formed deoxynucleoside triphosphate can then be incorporated into the growing primer. This effect was especially obvious when the synthetic template E3PN as sequenced. When the first correct nucleotide (dCTP) is added some of the non-degraded dTDP is converted to dTTP. After dCMP has been incorporated some of the formed dTTP can be incorporated. This out-of-phase obtained DNA can be further extended when dGTP is added. This is clearly shown when the out-of-phase DNA has reached the position where two A should be incorporated. The false signal is now stronger. The following double T and C also give stronger signals whereas the next single A gives a lower signal. In FIG. **5**, DNA sequencing of 20 bases of a 160-base-long self-primed single-stranded PCR product is shown. The obtained sequence was confirmed by semiautomatic solid-phase Sanger sequencing (data not shown). The main reason for the sequencing to come out of phase is a combination of slow degration of deoxynucleoside diphosphates (at least some of the dNDPs) by the potato apyrase (Liebecq, C. Lallemand A, and Deguldre-Guillaume, M. J. (1963) Bull. Soc. Chim. Biol. 45, 573–594) and the deoxynucleoside diphosphate kinase contamination in the ATP sulfurylase preparation obtained from Sigma. It is possible to overcome this problem by using a pure preparation of ATP sulfurylase, or by using more efficient dNDP degrading enzymes (Doremus, H. D. and Blevins, D. G. (1988) Plant Physiol. 87(1), 41–45). Even if a pure preparation of ATP sulfurylase is used it could be an advantage to use combinations of nucleotide degrading enzymes (NTPase, NDPase, NMPase) to increase the rate of the degradation process and to decrease the thermodynamic equilibrium concentration of dNTPs. In addition, it could be an advantage to use an enzyme with low Km for dNTPs such as the Pig Pancreas nucleoside triphosphate diphosphohydrolase (Le Bel, D., Piriet, G. G. Phaneuf, S., St-Jean, P., Laliberte, J. F. and Beudoin, A. R. (1980) J. Biol. Chem. 255, 1227–1233; Laliberte, J. F. St-Jean, P. and Beudoin, R. (1982) J. Biol Chem. 257, 3869–3871).

EXAMPLE 2

PyroSequencing" on a PCR Product

The biotinylated PCR products were immobilized onto streptavidin-coated super paramagnetic beads Dynabeads™ M280-Streptavidin (Dynal). Elution of single-stranded DNA and hybridization of sequencing primer (JA 80 5'-GATGGAAACCAAAAATGATAGG-3') SEQ ID NO:9 was carried out as described earlier (T. Hultman, M. Murby, S. Ståhl, E. Hornes, M. Uhlén, Nucleic Acids Res. 18: 5107 (1990)). The hybridized template/primer were incubated with Sequenase 2.0 DNA polymerase (Amersham). The sequencing procedure was carried out by stepwise elongation of the primer-strand upon sequential addition of the different deoxynucleoside triphosphates (Pharmacia Biotech), and simultaneous degradation of nucleotides by apyrase. The apyrase was grade VI, high ATPase/ADPase ratio (nucleoside 51-triphosphatase and nucleoside 5'-diphosphatase; EC 3.61.5) (Sigma Chemical Co.). The sequencing reaction was performed at room temperature and started by adding 0.6 nmol of one of the deoxynucleotides (Pharmacia Biotech). The PPi released due to nucleotide incorporation was detected as described earlier (see e.g. Example 1). The JA80 was synthesized by phosphoramidite chemistry (Interactiva). The sequencing data obtained from the PyroSequencing method was confirmed by semi-automated solid-phase Sanger sequencing according to Hultman et al. (T. Hultman, M. Murby, S. Ståhl, E. Hornes, M. Uhlén, Nucleic Acids Res. 18: 5107 (1990)). The reaction was carried out at room temperature. The results are shown in FIG. **6**.

SEQUENCE LISTING

<160> NUMBER OF SEQ ID NOS: 9

<210> SEQ ID NO 1
<211> LENGTH: 35
<212> TYPE: DNA
<213> ORGANISM: Artificial Sequence
<220> FEATURE:
<223> OTHER INFORMATION: Oligonucleotide (PEBE25)

-continued

<400> SEQUENCE: 1

gcaacgtcgc cacacacaac atacgagccg gaagg 35

<210> SEQ ID NO 2
<211> LENGTH: 23
<212> TYPE: DNA
<213> ORGANISM: Artificial Sequence
<220> FEATURE:
<223> OTHER INFORMATION: Oligonucleotide (RIT 27)

<400> SEQUENCE: 2

gcttccggct cgtatgttgt gtg 23

<210> SEQ ID NO 3
<211> LENGTH: 35
<212> TYPE: DNA
<213> ORGANISM: Artificial Sequence
<220> FEATURE:
<223> OTHER INFORMATION: Oligonucleotide (E3PN)

<400> SEQUENCE: 3

gctggaattc gtcagactgg ccgtcgtttt acaac 35

<210> SEQ ID NO 4
<211> LENGTH: 17
<212> TYPE: DNA
<213> ORGANISM: Artificial Sequence
<220> FEATURE:
<223> OTHER INFORMATION: Oligonucleotide (NUSPT)

<400> SEQUENCE: 4

gtaaaacgac ggccagt 17

<210> SEQ ID NO 5
<211> LENGTH: 51
<212> TYPE: DNA
<213> ORGANISM: Artificial Sequence
<220> FEATURE:
<223> OTHER INFORMATION: Oligonucleotide (RIT 203)

<400> SEQUENCE: 5

agcttgggtt cgaggagatc ttccgggtta cggcggaaga tctcctcgag g 51

<210> SEQ ID NO 6
<211> LENGTH: 51
<212> TYPE: DNA
<213> ORGANISM: Artificial Sequence
<220> FEATURE:
<223> OTHER INFORMATION: Oligonucleotide (RIT 204)

<400> SEQUENCE: 6

agctcctcga ggagatcttc cgccgtaacc cggaagatct cctcgaaccc a 51

<210> SEQ ID NO 7
<211> LENGTH: 25
<212> TYPE: DNA
<213> ORGANISM: Artificial Sequence
<220> FEATURE:
<223> OTHER INFORMATION: Oligonucleotide (ROMO 205S)

<400> SEQUENCE: 7

cgaggagatc ttccgggtta cggcg 25

<210> SEQ ID NO 8

-continued

<211> LENGTH: 25
<212> TYPE: DNA
<213> ORGANISM: Artificial Sequence
<220> FEATURE:
<223> OTHER INFORMATION: Oligonucleotide (ROMO 205B)
<221> NAME/KEY: modified_base
<222> LOCATION: (1)...(1)
<223> OTHER INFORMATION: biotin

<400> SEQUENCE: 8

cgaggagatc ttccgggtta cggcg 25

<210> SEQ ID NO 9
<211> LENGTH: 22
<212> TYPE: DNA
<213> ORGANISM: Artificial Sequence
<220> FEATURE:
<223> OTHER INFORMATION: Oligonucleotide (JA 80)

<400> SEQUENCE: 9

gatggaaacc aaaaatgata gg 22

What is claimed is:

1. A method of identifying a base at a target position in a sample DNA sequence comprising providing a sample DNA sequence and an extension primer, which hybridizes to the sample DNA immediately adjacent to the target position and subjecting the sample DNA and extension primer to a polymerase reaction in the presence of a deoxynucleotide or dideoxynucleotide whereby the deoxynucleotide or dideoxynucleotide will only become incorporated and release pyrophosphate (PPi) if it is complementary to the base in the target position, and detecting any release of PPi enzymically, different deoxynucleotides or dideoxynucleotides being added either to separate aliquots of sample-primer mixture or successively to the sample-primer mixture and subjected to the polymerase reaction to indicate which deoxynucleotide or dideoxynucleotide is incorporated, wherein a nucleotide-degrading enzyme is included during the polymerase reaction step, such that unincorporated nucleotides are degraded, and whereby any release of PPi is indicative of incorporation of deoxynucleotide or dideoxynucleotide and the identification of a base complementary thereto.

2. A method as claimed in claim **1**, wherein the nucleotide-degrading enzyme is apyrase.

3. A method as claimed in claim **1**, wherein a mixture of nucleotide-degrading enzymes is used having nucleoside triphosphatase, nucleoside diphosphatase and nucleoside monophosphatase activity.

4. A method as claimed in claim **1**, wherein the nucleotide-degrading enzyme is immobilised on a solid support.

5. A method as claimed in claim **4**, wherein said immobilised nucleotide-degrading enzyme is added after nucleotide incorporation by the polymerase has taken place, and then removed prior to a subsequent nucleotide incorporation reaction step.

6. A method as claimed in claim **1**, wherein PPi release is detected using the Enzymatic Luminometric Inorganic Pyrophosphate Detection Assay (ELIDA).

7. A method as claimed in claim **1**, wherein the PPi detection enzymes are included in the polymerase reaction step and the polymerase reaction and PPi release detection steps are performed substantially simultaneously.

8. A method as claimed in claim **1**, wherein in the polymerase reaction a dATP or ddATP analogue is used which is capable of acting as a substrate for a polymerase but incapable of acting as a substrate for a PPi detection enzyme.

9. A method as claimed in claim **8**, wherein the dATP analogue is deoxyadenosine α-thiotriphosphate (dATPαS).

10. A method as claimed in claim **1**, further comprising the use of the α-thio analogues of dCTP, dGTP and dTTP.

11. A method as claimed in claim **1**, wherein the sample DNA is immobilised or provided with means for attachment to a solid support.

12. A method as claimed in claim **1**, wherein the sample DNA is first amplified.

13. A method as claimed in claim **1**, wherein the extension primer contains a loop and anneals back on itself and the 3' end of the sample DNA.

14. A method as claimed in claim **1**, wherein an exonuclease deficient (exo^-) high fidelity polymerase is used.

15. A method as claimed in claim **1**, for identification of a base in a single target position in a DNA sequence comprising subjecting the sample DNA to amplification; immobilizing the amplified DNA and then subjecting the immobilized DNA to strand separation, removing the non-immobilized strand and providing an extension primer, which hybridizes to the immobilized DNA immediately adjacent to the target position; subjecting each of four aliquots of the immobilized single stranded DNA to a polymerase reaction in the presence of a dideoxynucleotide, each aliquot using a different dideoxynucleotide whereby only the dideoxynucleotide complementary to the base in the target position becomes incorporated; subjecting the four aliquots to extension in the presence of all four deoxynucleotides, whereby in each aliquot the DNA which has not reacted with the dideoxynucleotide is extended to form double stranded DNA while the dideoxy-blocked DNA remains as single stranded DNA; followed by identifying the double stranded and/or single stranded DNA to indicate which dideoxynucleotide was incorporated and hence which base was present in the target position.

16. A kit for use in a method as defined in claim **1**, comprising:

(a) a polymerase;

(b) detection enzyme means for identifying pyrophosphate release;

(c) a mixture of nucleotide-degrading enzymes;

(d) deoxynucleotides, or optionally deoxynucleotide analogues, optionally including, in place of dATP, a dATP analogue which is capable of acting as a substrate for a polymerase but incapable of acting as a substrate for a said PPi-detection enzyme; and

(e) optionally a test specific primer which hybridises to sample DNA so that the target position is directly adjacent to the 3' end of the primer; and

(f) optionally dideoxynucleotides, or optionally dideoxynucleotide analogues, optionally ddATP being replaced by a ddATP analogue which is capable of acting as a substrate for a polymerase but incapable of acting as a substrate for a said PPi-detection enzyme.

17. A method of identifying a base at a target position in a sample DNA sequence comprising arranging a multiplicity of DNA sequences in array format on a solid surface, providing to each sample an extension primer, which hybridizes to the sample DNA immediately adjacent to the target position and subjecting the sample DNA and extension primer to a polymerase reaction in the presence of a deoxynucleotide or dideoxynucleotide whereby the deoxynucleotide or dideoxynucleotide will only become incorporated and release pyrophosphate (PPi) if it is complementary to the base in the target position and detecting any release of PPi enzymically, different deoxynucleotides or dideoxynucleotides being added either to separate aliquots of sample-primer mixture or successively to the same sample-primer mixture and subjected to the polymerase reaction to indicate which deoxynucleotide or dideoxynucleotide is incorporated whereby a nucleotide-degrading enzyme is included during the polymerase reaction step such that unincorporated nucleotides are degraded, and whereby any release of PPi is indicative of incorporation of deoxynucleotide or dideoxynucleotide and the identification of a base complementary thereto.

* * * * *

US005786146A

United States Patent [19]

Herman et al.

[11] Patent Number: 5,786,146

[45] Date of Patent: Jul. 28, 1998

[54] **METHOD OF DETECTION OF METHYLATED NUCLEIC ACID USING AGENTS WHICH MODIFY UNMETHYLATED CYTOSINE AND DISTINGUISHING MODIFIED METHYLATED AND NON-METHYLATED NUCLEIC ACIDS**

[75] Inventors: **James G. Herman**, Timonium; **Stephen B. Baylin**, Baltimore, both of Md.

[73] Assignee: **The Johns Hopkins University School of Medicine**, Baltimore, Md.

[21] Appl. No.: **656,716**

[22] Filed: **Jun. 3, 1996**

[51] **Int. Cl.**[6] **C12Q 1/68**; C12P 19/34; C07H 21/04

[52] **U.S. Cl.** **435/6**; 435/91.2; 536/24.31; 536/24.33

[58] **Field of Search** 435/6, 91.2; 536/22.1, 536/23.1, 23.5, 24.3, 24.33

[56] **References Cited**

U.S. PATENT DOCUMENTS

5,324,634	6/1994	Zucker	435/7.23
5,595,885	1/1997	Stetler-Stevenson et al.	435/69.2

OTHER PUBLICATIONS

Herman et al., Proc. Natl. Acad. Sci. USA 93, 9821–9826 (Sep. 1996).

Stetler–Stevenson et al., "Tissue Inhibitor of Metalloproteinases–2 (TIMP–2) mRNA Expression in Tumor Cell Lines and Human Tumor Tissues," *J. Biol. Chem.* Aug. 15, 1990, vol. 265, No. 23, pp. 13933–13938.

Frommer et al., "A genomic sequencing protocol that yields a positive display of 5–methylcyltosine residues in individual DNA strands," *Proc. Natl. Acad. Sci. USA*, Mar. 1992, vol. 89, pp. 1827–1831.

Gonzales–Zulueta, et al., Methylation of the 5'CpG Island of the p16/CDKNA2 Tumor Suppressor Gene in Normal and Transformed Human Tissues Correlates with Gene Silencing, Cancer Research 55:4531–4535, Oct. 1995.

Zuccotti, et al., Polymerase Chain Reaction for the Detection of Methylation of a Specific CpG Site in the G6pd Gene of Mouse Embryos, Methods in Enzymology 225:557–567, 1993.

Clark, et al., *High sensitivity mapping of methylated cytosines, Nucleic Acid Research*, 22(15):2990, 1994.

Frommer, et al., *A genomic sequencing protocol that yields a positive display of 5–methylcytosine residues in individual DNA strands, Proc. Natl. Acad. Sci. USA*, 89:1827, Mar. 1992.

Graff, et al., *E–Cadherin Expression is Silenced by DNA Hypermethylation in Human Breast and Prostate Carcinomas, Cancer Research*, 55:5195, Nov. 15, 1955.

Herman, et al., *Inactivation of the CDKN2/p16/MTS1 Gene is Frequently Associated with Aberrant DNA Methylation in All Common Human Cancers, Cancer Research*, 55:4525, Oct. 15, 1995.

Lowe, et al., *A computer program for selection of oligonucleotide primers for polymerase chain reactions, Nucleic Acids research*, 18(7):1757, Mar. 2, 1990.

Myohanene, et al., *Automated fluorescent genomic sequencing as applied to the methylation analysis of the human orthine decarboxylase gene, DNA Sequence–The Journal of Sequencing and Mapping*, 5:1, 1994.

Park, et al., *CpG island Promoter region Methylation Patterns of the Inactive–X–Chromosome Hypoxanthine Phosphoribosyltransferase (Hprt) Gene, Molecular and Celluar Biology*, 14(12):7975, Dec. 1994.

Raizis, et al., *A Bisulfite Method of 5–Methylcytosine Mapping That Minimizes Template Degradation, Analytical Biochemistry*, 226:161, 1995.

Reeben, et al., *Sequencing of the rat light neurofilament promoter reveals difference in ethylation between expressing and non–expressing cell lines, but not tissues, Gene*, 157:325, 1995.

Tasheva and Roufa, *Deoxycytidine Methylation and the Origin of Spontaneous transition Mutations in Mammalian Cells, Samatic Cell and Molecular genetics*, 19(3):275, 1993.

Primary Examiner—Kenneth R. Horlick

Attorney, Agent, or Firm—Fish & Richardson, P.C.

[57] **ABSTRACT**

The present invention provides a method of PCR, methylation specific PCR (MSP), for rapid identification of DNA methylation patterns in a CpG-containing nucleic acid. MSP uses agents to modify unmethylated cytosine in a nucleic acid of interest, and then uses the PCR reaction to amplify the CpG-containing nucleic acid in the specimen by means of CpG-specific oligonucleotide primers. The oligonecleotide primers distinguish between modified methylated and nonmethylated nucleic acid. Kits utilizing MSP for the detection of methylated CpG-containing nucleic acids are also provided.

27 Claims, 3 Drawing Sheets

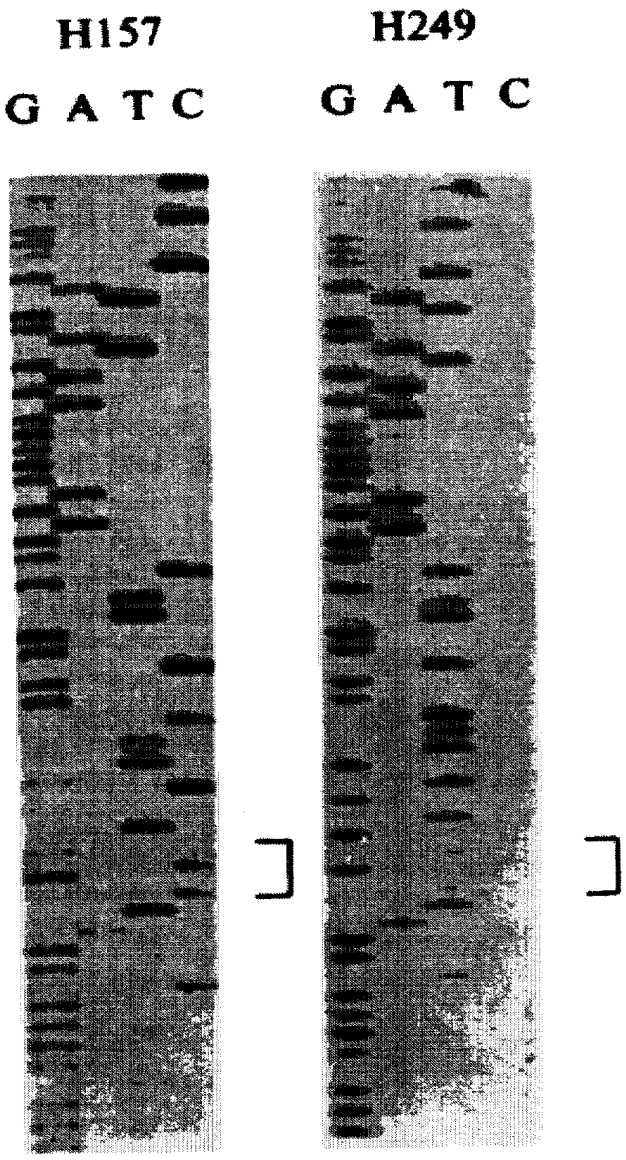


FIG. 1

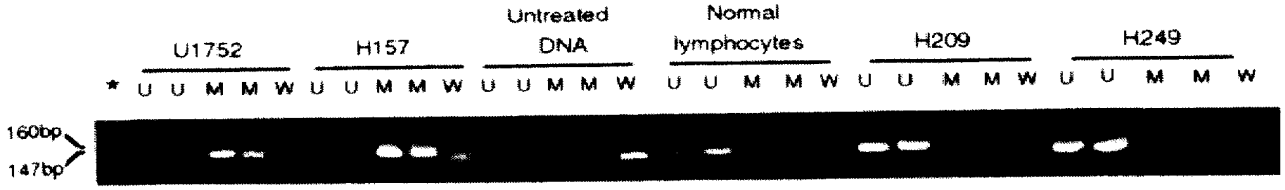


FIG. 2A

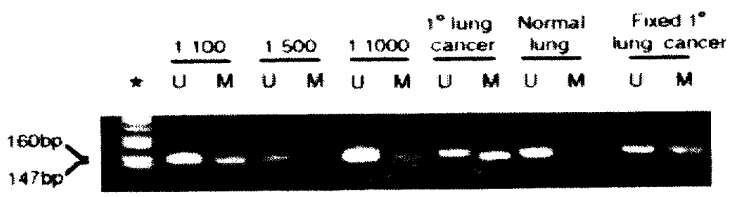


FIG. 2B

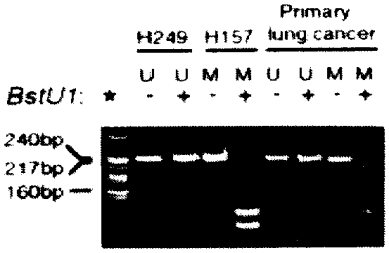


FIG. 2D

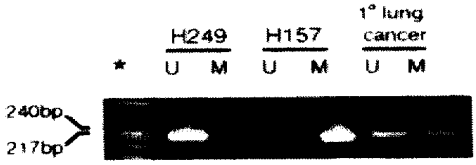


FIG. 2C

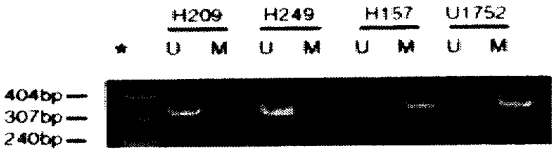


FIG. 2E

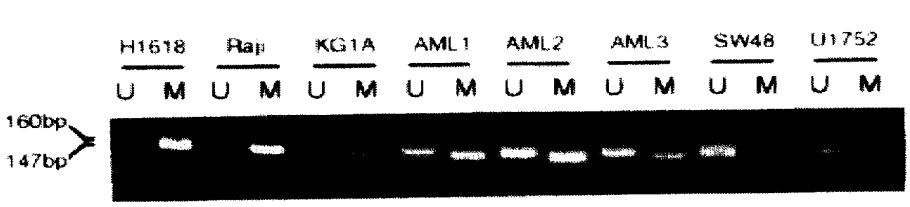


FIG. 3A

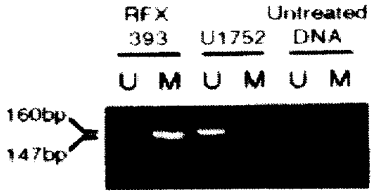


FIG. 3C

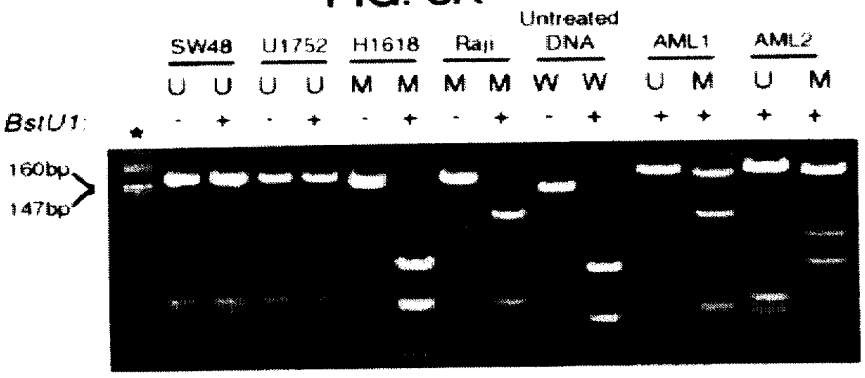


FIG. 3B

FIG. 3D

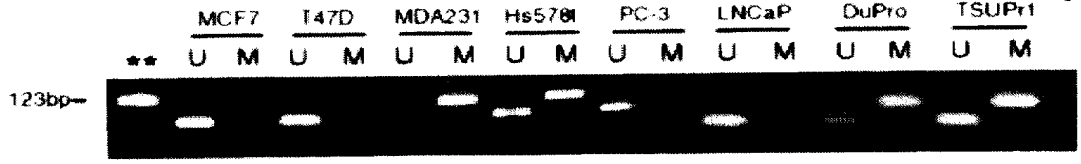


FIG. 3E

METHOD OF DETECTION OF METHYLATED NUCLEIC ACID USING AGENTS WHICH MODIFY UNMETHYLATED CYTOSINE AND DISTINGUISHING MODIFIED METHYLATED AND NON-METHYLATED NUCLEIC ACIDS

This invention was made with government support under Grant Nos. CA43318 and CA54396 awarded by the National Institutes of Health. The government has certain rights in this invention.

FIELD OF THE INVENTION

The present invention relates generally to regulation of gene expression, and more specifically to a method of determining the DNA methylation status of CpG sites in a given locus.

BACKGROUND OF THE INVENTION

In higher order eukaryotes DNA is methylated only at cytosines located 5' to guanosine in the CpG dinucleotide. This modification has important regulatory effects on gene expression, especially when involving CpG rich areas, known as CpG islands, located in the promoter regions of many genes. While almost all gene-associated islands are protected from methylation on autosomal chromosomes, extensive methylation of CpG islands has been associated with transcriptional inactivation of selected imprinted genes and genes on the inactive X -chromosome of females. Abberant methylation of normally unmethylated CpG islands has been described as a frequent event in immortalized and transformed cells, and has been associated with transcriptional inactivation of defined tumor suppressor genes in human cancers.

Human cancer cells typically contain somatically altered genomes, characterized by mutation, amplification, or deletion of critical genes. In addition, the DNA template from human cancer cells often displays somatic changes in DNA methylation (E. R. Fearon, et al., *Cell*, 61:759, 1990; P. A. Jones, et al., *Cancer Res.*, 46:461, 1986; R. Holliday, *Science*, 238:163, 1987; A. De Bustros, et aL, *Proc. Natl. Acad. Sci., USA*, 85:5693, 1988); P. A. Jones, et al., *Adv. Cancer Res.*, 54:1, 1990; S. B. Baylin, et al., *Cancer Cells*, 3 :383, 1991; M. Makos, et al, *Proc. Natl. Acad. Sci., USA*, 89:1929, 1992; N. Ohtani-Fujita, et al., *Oncogene*, 8:1063, 1993). However, the precise role of abnormal DNA methylation in human tumorigenesis has not been established. DNA methylases transfer methyl groups from the universal methyl donor S-adenosyl methionine to specific sites on the DNA. Several biological functions have been attributed to the methylated bases in DNA. The most established biological function is the protection of the DNA from digestion by cognate restriction enzymes. The restriction modification phenomenon has, so far, been observed only in bacteria. Mammalian cells, however, possess a different methylase that exclusively methylates cytosine residues on the DNA, that are 5' neighbors of guanine (CpG). This methylation has been shown by several lines of evidence to play a role in gene activity, cell differentiation, tumorigenesis, X-chromosome inactivation, genomic imprinting and other major biological processes (Razin, A., H., and Riggs, R. D. eds. in DNA Methylation Biochemistry and Biological Significance, Springer-Verlag, N.Y., 1984).

A CpG rich region, or "CpG island", has recently been identified at 17p13.3, which is aberrantly hypermethylated

in multiple common types of human cancers (Makos, M., et al., *Proc. Natl. Acad. Sci. USA*, 89:1929, 1992; Makos, M., et al, *Cancer Res.*, 53:2715, 1993; Makos, M., et al, *Cancer Res.* 53:2719, 1993). This hypermethylation coincides with timing and frequency of 17p losses and p53 mutations in brain, colon, and renal cancers. Silenced gene transcription associated with hypermethylation of the normally unmethylated promoter region CpG islands has been implicated as an alternative mechanism to mutations of coding regions for inactivation of tumor suppressor genes (Baylin, S. B., et al, *Cancer Cells*, 3:383, 1991; Jones, P. A. and Buckley, J. D., *Adv. Cancer Res.*, 54:1–23, 1990). This change has now been associated with the loss of expression of VHL, a renal cancer tumor suppressor gene on 3p (J. G. Herman, et al, *Proc. Natl. Acad. Sci. USA*, 91:9700–9704, 1994), the estrogen receptor gene on 6q (Ottaviano, Y. L., et al., *Cancer Res.*, 54:2552, 1994) and the H19 gene on 11p (Steenman, M. J. C., et al, *Nature Genetics*, 7:433, 1994).

In eukaryotic cells, methylation of cytosine residues that are immediately 5' to a guanosine, occurs predominantly in CG poor regions (Bird, A., *Nature*, 321:209, 1986). In contrast, discrete regions of CG dinucleotides called CpG islands remain unmethylated in normal cells, except during X-chromosome inactivation (Migeon, et al., supra) and parental specific imprinting (Li, et al, *Nature*, 366:362, 1993) where methylation of 5' regulatory regions can lead to transcriptional repression. De novo methylation of the Rb gene has been demonstrated in a small fraction of retinoblastomas (Sakai, et al, *Am. J. Hum. Genet.*, 48:880, 1991), and recently, a more detailed analysis of the VHL gene showed aberrant methylation in a subset of sporadic renal cell carcinomas (Herman, et al., *Proc. Natl. Acad. Sci., U.S.A.*, 91:9700, 1994). Expression of a tumor suppressor gene can also be abolished by de novo DNA methylation of a normally unmethylated 5' CpG island (Issa, et al, *Nature Genet.*, 7:536, 1994; Herman, et al., supra; Merlo, et al., *Nature Med.*, 1:686, 1995; Herman, et al., *Cancer Res.*, 56:722, 1996; Graff, et al, *Cancer Res.*, 55:5195, 1995; Herman, et al, *Cancer Res.*, 55:4525, 1995).

Most of the methods developed to date for detection of methylated cytosine depend upon cleavage of the phosphodiester bond alongside cytosine residues, using either methylation-sensitive restriction enzymes or reactive chemicals such as hydrazine which differentiate between cytosine and its 5-methyl derivative. The use of methylation-sensitive enzymes suffers from the disadvantage that it is not of general applicability, since only a limited proportion of potentially methylated sites in the genome can be analyzed. Genomic sequencing protocols which identify a 5-MeC residue in genomic DNA as a site that is not cleaved by any of the Maxam Gilbert sequencing reactions, are a substantial improvement on the original genomic sequencing method, but still suffer disadvantages such as the requirement for large amount of genomic DNA and the difficulty in detecting a gap in a sequencing ladder which may contain bands of varying intensity.

Mapping of methylated regions in DNA has relied primarily on Southern hybridization approaches, based on the inability of methylation-sensitive restriction enzymes to cleave sequences which contain one or more methylated CpG sites. This method provides an assessment of the overall methylation status of CpG islands, including some quantitative analysis, but is relatively insensitive, requires large amounts of high molecular weight DNA and can only provide information about those CpG sites found within sequences recognized by methylation-sensitive restriction enzymes. A more sensitive method of detecting methylation

patterns combines the use of methylation-sensitive enzymes and the polymerase chain reaction (PCR). After digestion of DNA with the enzyme, PCR will amplify from primers flanking the restriction site only if DNA cleavage was prevented by methylation. Like Southern-based approaches, this method can only monitor CpG methylation in methylation-sensitive restriction sites. Moreover, the restriction of unmethylated DNA must be complete, since any uncleaved DNA will be amplified by PCR yielding a false positive result for methylation. This approach has been useful in studying samples where a high percentage of alleles of interest are methylated, such as the study of imprinted genes and X-chromosome inactivated genes. However, difficulties in distinguishing between incomplete restriction and low numbers of methylated alleles make this approach unreliable for detection of tumor suppressor gene hypermethylation in small samples where methylated alleles represent a small fraction of the population.

Another method that avoids the use of restriction endonucleases utilizes bisulfite treatment of DNA to convert all unmethylated cytosines to uracil. The altered DNA is amplified and sequenced to show the methylation status of all CpG sites. However, this method is technically difficult, labor intensive and without cloning amplified products, it is less sensitive than Southern analysis, requiring approximately 10% of the alleles to be methylated for detection.

Identification of the earliest genetic changes in tumorigenesis is a major focus in molecular cancer research. Diagnostic approaches based on identification of these changes are likely to allow implementation of early detection strategies and novel therapeutic approaches targeting these early changes might lead to more effective cancer treatment.

SUMMARY OF THE INVENTION

The precise mapping of DNA methylation patterns in CpG islands has become essential for understanding diverse biological processes such as the regulation of imprinted genes, X-chromosome inactivation, and tumor suppressor gene silencing in human cancer. The present invention provides a method for rapid assessment of the methylation status of any group of CpG sites within a CpG island, independent of the use of methylation-sensitive restriction enzymes.

The method of the invention includes modification of DNA by sodium bisulfite or a comparable agent which converts all unmethylated but not methylated cytosines to uracil, and subsequent amplification with primers specific for methylated versus unmethylated DNA. This method of "methylation specific PCR" or MSP, requires only small amounts of DNA, is sensitive to 0.1% of methylated alleles of a given CpG island locus, and can be performed on DNA extracted from paraffin-embedded samples, for example. MSP eliminates the false positive results inherent to previous PCR-based approaches which relied on differential restriction enzyme cleavage to distinguish methylated from unmethylated DNA.

In a particular aspect of the invention, MSP is useful for identifying promoter region hypermethylation changes associated with transcriptional inactivation in tumor suppressor genes, for example, p16, p15, E-cadherin and VHL, in human neoplasia.

BRIEF DESCRIPTION OF THE DRAWINGS

FIG. 1 shows genomic sequencing of p16. The sequence shown has the most 5' region at the bottom of the gel, beginning at +175 in relation to a major transcriptional start site (Hara, et al., *Mol. Cell Biol.* 16:859, 1996). All cytosines in the unmethylated cell line H249 have been converted to thymidine, while all C's in CpG dinucleotides in the methylated cell H157 remains as C, indicating methylation.] enclosed a BstUI site which is at −59 in relation to the transnational start site in Genbank sequence U12818 (Hussussian, et al. *Nat. Genet.*, 8:15, 1994), but which is incorrectly identified as CGCA in sequence X94154 (Hara, et al., supra). This CGCG site represents the 3' location of the sense primer used for p16 MSP.

FIGS. 2A–2E show polyacrylamide gels with the Methylation Specific PCR products of p16. Primer sets used for amplification are designated as unmethylated (U), methylated (M), or unmodified/wild-type (W). * designates the molecular weight marker pBR322-MspI digest. Panel A shows amplification of bisulfite-treated DNA from cancer cell lines and normal lymphocytes, and untreated DNA (from cell line H249). Panel B shows mixing of various amount of H157 DNA with 1 μg of H249 DNA prior to bisulfite treatment to assess the detection sensitivity of MSP for methylated alleles. Modified DNA from a primary lung cancer sample and normal lung are also shown. Panel C shows amplification with the p16-U2 (U) primers, and p16-M2 (M) described in Table 1. Panel D shows the amplified p16 products of panel C restricted with BstUI(+) or not restricted (−). Panel E shows results of testing for regional methylation of CpG islands with MSP, using sense primers p16-U2 (U) and p16-M2 (M), which are methylation specific, and an antisense primer which is not methylation specific.

FIGS. 3A–3E show polyacrylaride gels of MSP products from analysis of several genes. Primer sets used for amplification are not designated as unmethylated (U), methylatecd (M), or unmodified/wild-type (W). * designates the molecular weight marker pBR322-MspI digest and ** designates the 123 bp molecular weight marker. All DNA samples were bisulfite treated except those designated untreated. Panel A shows the results from MSP for p15. Panel B shows the p15 products restricted with BstUI (+) or not restricted (−). Panel C shows the products of MSP for VHL. Panel D shows the VHL products restricted with BstUI(+) or not restricted (−). Panel E shows the products of MSP for E-cadherin.

DESCRIPTION OF THE PREFERRED EMBODIMENTS

The present invention provides methylation specific PCR (MSP) for identification of DNA methylation patterns. MSP uses the PCR reaction itself to distinguish between methylated and unmethylated DNA, which adds an improved sensitivity of methylation detection.

Unlike previous genomic sequencing methods for methylation identification which utilizes amplification primers which are specifically designed to avoid the CpG sequences, MSP primers are specifically designed to recognize CpG sites to take advantage of the differences in methylation to amplify specific products to be identified by the invention assay.

As illustrated in the Examples below, MSP provides significant advantages over previous PCR and other methods used for assaying methylation. MSP is markedly more sensitive than Southern analyses, facilitating detection of low numbers of methylated alleles and the study of DNA from small samples. MSP allows the study of paraffin-embedded materials, which could not previously be analyzed by Southern analysis. MSP also allows examination of

all CpG sites, not just those within sequences recognized by methylation-sensitive restriction enzymes. This markedly increases the number of such sites which can be assessed and will allow rapid, fine mapping of methylation patterns throughout CpG rich regions. MSP also eliminates the frequent false positive results due to partial digestion of methylation-sensitive enzymes inherent in previous PCR methods for detecting methylation. Furthermore, with MSP, simultaneous detection of unmethylated and methylated products in a single sample confirms the integrity of DNA as a template for PCR and allows a semi-quantitative assessment of allele types which correlates with results of Southern analysis. Finally, the ability to validate the amplified product by differential restriction patterns is an additional advantage.

The only technique that can provide more direct analysis than MSP for most CpG sites within a defined region is genomic sequencing. However, MSP can provide similar information and has the following advantages. First, MSP is much simpler and requires less than genomic sequencing, with a typical PCR and gel analysis taking 4–6 hours. In contrast, genomic sequencing, amplification, cloning, and subsequent sequencing may take days. MSP also avoids the use of expensive sequencing reagents and the use of radioactivity. Both of these factors make MSP better suited for the analysis of large numbers of samples. Third, the use of PCR as the step to distinguish methylated from unmethylated DNA in MSP allows for significant increase in the sensitivity of methylation detection. For example, if cloning is not used prior to genomic sequencing of the DNA, less than 10% methylated DNA in a background of unmethylated DNA cannot be seen (Myohanen, et al, supra). The use of PCR and cloning does allow sensitive detection of methylation patterns in very small amounts of DNA by genomic sequencing (Frommer, et al, *Proc. Natl Acad. Sci. USA*, 89:1827, 1992; Clark, et al, *Nucleic Acids Research*, 22:2990, 1994). However, this means in practice that it would require sequencing analysis of 10 clones to detect 10% methylation, 100 clones to detect 1% methylation, and to reach the level of sensitivity we have demonstrated with MSP (1:1000), one would have to sequence 1000 individual clones.

In a first embodiment, the invention provides a method for detecting a methylated CpG-containing nucleic acid, the method including contacting a nucleic acid-containing specimen with an agent that modifies unmethylated cytosine; amplifying the CpG-containing nucleic acid in the specimen by means of CpG-specific oligonucleotide primers; and detecting the methylated nucleic acid. It is understood that while the amplification step is optional, it is desirable in the preferred method of the invention.

The term "modifies" as used herein means the conversion of an unmethylated cytosine to another nucleotide which will distinguish the unmethylated from the methylated cytosine. Preferably, the agent modifies unmethylated cytosine to uracil. Preferably, the agent used for modifying unmethylated cytosine is sodium bisulfite, however, other agents that similarly modify unmethylated cytosine, but not methylated cytosine can also be used in the method of the invention. Sodium bisulfite ($NaHSO_3$) reacts readily with the 5,6-double bond of cytosine, but poorly with methylated cytosine. Cytosine reacts with the bisulfite ion to form a sulfonated cytosine reaction intermediate which is susceptible to deamination, giving rise to a sulfonated uracil. The sulfonate group can be removed under alkaline conditions, resulting in the formation of uracil. Uracil is recognized as a thymine by Taq polymerase and therefore upon PCR, the resultant product contains cytosine only at the position where 5-methylcytosine occurs in the starting template DNA.

The primers used in the invention for amplification of the CpG-containing nucleic acid in the specimen, after bisulfite modification, specifically distinguish between untreated DNA, methylated, and non-methylated DNA. MSP primers for the non-methylated DNA preferably have a T in the 3' CG pair to distinguish it from the C retained in methylated DNA, and the compliment is designed for the antisense primer. MSP primers usually contain relatively few Cs or Gs in the sequence since the Cs will be absent in the sense primer and the Gs absent in the antisense primer (C becomes modified to U (uracil) which is amplified as T (thymidine) in the amplification product).

The primers of the invention embrace oligonucleotides of sufficient length and appropriate sequence so as to provide specific initiation of polymerization on a significant number of nucleic acids in the polymorphic locus. Specifically, the term "primer" as used herein refers to a sequence comprising two or more deoxyribonucleotides or ribonucleotides, preferably more than three, and most preferably more than 8, which sequence is capable of initiating synthesis of a primer extension product, which is substantially complementary to a polymorphic locus strand. Environmental conditions conducive to synthesis include the presence of nucleoside triphosphates and an agent for polymerization, such as DNA polymerase, and a suitable temperature and pH. The primer is preferably single stranded for maximum efficiency in amplification, but may be double stranded. If double stranded, the primer is first treated to separate its strands before being used to prepare extension products. Preferably, the primer is an oligodeoxy ribonucleo-tide. The primer must be sufficiently long to prime the synthesis of extension products in the presence of the inducing agent for polymerization. The exact length of primer will depend on many factors, including temperature, buffer, and nucleotide composition. The oligonucleotide primer typically contains 12–20 or more nucleotides, although it may contain fewer nucleotides.

Primers of the invention are designed to be "substantially" complementary to each strand of the genomic locus to be amplified and include the appropriate G or C nucleotides as discussed above. This means that the primers must be sufficiently complementary to hybridize with their respective strands under conditions which allow the agent for polymerization to perform. In other words, the primers should have sufficient complementarity with the 5' and 3' flanking sequences to hybridize therewith and permit amplification of the genomic locus.

Oligonucleotide primers of the invention are employed in the amplification process which is an enzymatic chain reaction that produces exponential quantities of target locus relative to the number of reaction steps involved. Typically, one primer is complementary to the negative (−) strand of the locus and the other is complementary to the positive (+) strand. Annealing the primers to denatured nucleic acid followed by extension with an enzyme, such as the large fragment of DNA Polymerase I (Klenow) and nucleotides, results in newly synthesized + and − strands containing the target locus sequence. Because these newly synthesized sequences are also templates, repeated cycles of denaturing, primer annealing, and extension results in exponential production of the region (i.e., the target locus sequence) defined by the primer. The product of the chain reaction is a discrete nucleic acid duplex with termini corresponding to the ends of the specific primers employed.

The oligonucleotide primers of the invention may be prepared using any suitable method, such as conventional phosphotriester and phosphodiester methods or automated

embodiments thereof. In one such automated embodiment, diethylphosphoramidites are used as starting materials and may be synthesized as described by Beaucage, et al. (*Tetrahedron Letters*, 22:1859–1862, 1981). One method for synthesizing oligonucleotides on a modified solid support is described in U.S. Pat. No. 4,458,066.

Any nucleic acid specimen, in purified or nonpurified form, can be utilized as the starting nucleic acid or acids, provided it contains, or is suspected of containing, the specific nucleic acid sequence containing the target locus (e.g., CpG). Thus, the process may employ, for example, DNA or RNA, including messenger RNA, wherein DNA or RNA may be single stranded or double stranded. In the event that RNA is to be used as a template, enzymes, and/or conditions optimal for reverse transcribing the template to DNA would be utilized. In addition, a DNA-RNA hybrid which contains one strand of each may be utilized. A mixture of nucleic acids may also be employed, or the nucleic acids produced in a previous amplification reaction herein, using the same or different primers may be so utilized. The specific nucleic acid sequence to be amplified, i.e., the target locus, may be a fraction of a larger molecule or can be present initially as a discrete molecule, so that the specific sequence constitutes the entire nucleic acid. It is not necessary that the sequence to be amplified be present initially in a pure form; it may be a minor fraction of a complex mixture, such as contained in whole human DNA.

The nucleic acid-containing specimen used for detection of methylated CpG may be from any source including brain, colon, urogenital, hematopoietic, thymus, testis, ovarian, uterine, prostate, breast, colon, lung and renal tissue and may be extracted by a variety of techniques such as that described by Maniatis, et al (*Molecular Cloning: A Laboratory Manual*, Cold Spring Harbor, N.Y., pp 280, 281, 1982).

If the extracted sample is impure (such as plasma, serum, or blood or a sample embedded in parrafin), it may be treated before amplification with an amount of a reagent effective to open the cells, fluids, tissues, or animal cell membranes of the sample, and to expose and/or separate the strand(s) of the nucleic acid(s). This lysing and nucleic acid denaturing step to expose and separate the strands will allow amplification to occur much more readily.

Where the target nucleic acid sequence of the sample contains two strands, it is necessary to separate the strands of the nucleic acid before it can be used as the template. Strand separation can be effected either as a separate step or simultaneously with the synthesis of the primer extension products. This strand separation can be accomplished using various suitable denaturing conditions, including physical, chemical, or enzymatic means, the word "denaturing" includes all such means. One physical method of separating nucleic acid strands involves heating the nucleic acid until it is denatured. Typical heat denaturation may involve temperatures ranging from about 80° to 105° C. for times ranging from about 1 to 10 minutes. Strand separation may also be induced by an enzyme from the class of enzymes known as helicases or by the enzyme RecA, which has helicase activity, and in the presence of riboATP, is known to denature DNA. The reaction conditions suitable for strand separation of nucleic acids with helicases are described by Kuhn Hoffmann-Berling (CSH-Quantitative Biology, 43:63, 1978) and techniques for using RecA are reviewed in C. Radding (*Ann. Rev. Genetics*, 16:405–437, 1982).

When complementary strands of nucleic acid or acids are separated, regardless of whether the nucleic acid was originally double or single stranded, the separated strands are ready to be used as a template for the synthesis of additional nucleic acid strands. This synthesis is performed under conditions allowing hybridization of primers to templates to occur. Generally synthesis occurs in a buffered aqueous solution, preferably at a pH of 7–9, most preferably about 8. Preferably, a molar excess (for genomic nucleic acid, usually about 10^8:1 primer:template) of the two oligonucleotide primers is added to the buffer containing the separated template strands. It is understood, however, that the amount of complementary strand may not be known if the process of the invention is used for diagnostic applications, so that the amount of primer relative to the amount of complementary strand cannot be determined with certainty. As a practical matter, however, the amount of primer added will generally be in molar excess over the amount of complementary strand (template) when the sequence to be amplified is contained in a mixture of complicated long-chain nucleic acid strands. A large molar excess is preferred to improve the efficiency of the process.

The deoxyribonucleoside triphosphates dATP, dCTP, dGTP, and dTTP are added to the synthesis mixture, either separately or together with the primers, in adequate amounts and the resulting solution is heated to about 90°–100° C. from about 1 to 10 minutes, preferably from 1 to 4 minutes. After this heating period, the solution is allowed to cool to room temperature, which is preferable for the primer hybridization. To the cooled mixture is added an appropriate agent for effecting the primer extension reaction (called herein "agent for polymerization"), and the reaction is allowed to occur under conditions known in the art. The agent for polymerization may also be added together with the other reagents if it is heat stable. This synthesis (or amplification) reaction may occur at room temperature up to a temperature above which the agent for polymerization no longer functions. Thus, for example, if DNA polymerase is used as the agent, the temperature is generally no greater than about 40° C. Most conveniently the reaction occurs at room temperature.

The agent for polymerization may be any compound or system which will function to accomplish the synthesis of primer extension products, including enzymes. Suitable enzymes for this purpose include, for example, *E. coli* DNA polymerase I, Klenow fragment of *E. coli* DNA polymerase I, T4 DNA polymerase, other available DNA polymerases, polymerase muteins, reverse transcriptase, and other enzymes, including heat-stable enzymes (i.e., those enzymes which perform primer extension after being subjected to temperatures sufficiently elevated to cause denaturation). Suitable enzymes will facilitate combination of the nucleotides in the proper manner to form the primer extension products which are complementary to each locus nucleic acid strand. Generally, the synthesis will be initiated at the 3' end of each primer and proceed in the 5' direction along the template strand, until synthesis terminates, producing molecules of different lengths. There may be agents for polymerization, however, which initiate synthesis at the 5' end and proceed in the other direction, using the same process as described above.

Preferably, the method of amplifying is by PCR, as described herein and as is commonly used by those of ordinary skill in the art. Alternative methods of amplification have been described and can also be employed as long as the methylated and non-methylated loci amplified by PCR using the primers of the invention is similarly amplified by the alternative means.

The amplified products are preferably identified as methylated or non-methylated by sequencing. Sequences amplified

by the methods of the invention can be further evaluated, detected, cloned, sequenced, and the like, either in solution or after binding to a solid support, by any method usually applied to the detection of a specific DNA sequence such as PCR, oligomer restriction (Saiki, et al., *BiolTechnology*, 3:1008–1012, 1985), allele-specific oligonucleotide (ASO) probe analysis (Conner, et al., *Proc. Natl. Acad. Sci. USA*, 80:278, 1983), oligonucleotide ligation assays (OLAs) (Landegren, et al., *Science*, 241:1077, 1988), and the like. Molecular techniques for DNA analysis have been reviewed (Landegren, et al., *Science*, 242:229–237, 1988).

Optionally, the methylation pattern of the nucleic acid can be confirmed by restriction enzyme digestion and Southern blot analysis. Examples of methylation sensitive restriction endonucleases which can be used to detect 5'CpG methylation include SmaI, SacII, EagI, MspI, HpaII, BstUI and BssHII, for example.

Exemplaiy target polynucleotide sequences to which the primer hybridizes have a sequence as listed below.

		SEQ ID NO.
Wild type p16	5'-GCGGTCCGCCCCACCCTCTG-3';	1
	5'-CCACGGCCGCGGCCCG-3';	2
Methylated p16-1*	5'-GCGATCCGCCCCACCCTCTAATAA-3';	3
	5'-TTACGGTCGCGGTTCGGGGTC-3';	4
Unmethylated p16-1	5'-ACAATCCACCCCACCCTCTAATAA-3';	5
	5'-TTATGGTTGTGGTTTGGGGTTG-3';	6
Methylated p16-2	5'-GCGATCCGCCCCACCCTCTAATAA-3'	7
	5'-CGGTCGGAGGTCGATTTAGGTGG-3'	8
Unmethylated p16-2	5'-ACAATCCACCCCACCCTCTAATAA-3';	9
	5'-TGGTTGGAGGTTGATTTAGGTGG-3';	10
Wild type p15	5'-TCTGGCCGCAGGGTGCG-3';	11
	5'-CCGGCCGCTCGGCCACT-3';	12
Methylated p15	5'-AACCGCAAAATACGAACGC-3';	13
	5'-TCGGTCGTTCGGTTATTGTACG-3';	14
Unmethylated p15	5'-AACCACAAAATACAAACACATCACA-3';	15
	5'-TTGGTTGTTTGGTTATTGTATGG-3';	16
Methylated VHL	5'-GCGTACGCAAAAAAATCCTCCA-3';	17
	5'-TTCGCGGCGTTCGGTTC-3';	18
Unmethylated VHL	5'-ACATACACAAAAAAATCCTCCAAC-3';	19
	5'-TTTGTGGTGTTTGGTTTGGG-3';	20
Methylated E-cadherin	5'-ACGCGATAACCCTCTAACCTAA-3';	21
	5'-GTCGGTAGGTGAATTTTTAGTTA-3';	22
Unmethylated E-cadherin	5'-ACAATAACCCTCTAACCTAAAATTA-3'; and	23
	5'-TGTGTTGTTTGATTGGTTGTG-3'.	24

Exemplary primer pairs included in the invention that hybridize to the above sequences include:

	SEQ ID NO:
5'-CAGAGGGTGGGGGCGGACCGC-3' and	26
5'-CGGGCCGCGGCCGTGG-3';	27
5'-TTATTAGAGGGTGGGGCGGATCGC-3' and	28
5'-GACCCCGAACCGCGACCGTAA-3';	29
5'-TTATTAGAGGGTGGGGTGGATTGT-3' and	30
5'-CAACCCCAAACCACAACCATAA-3';	31
5'-TTATTAGAGGGTGGGGCGGATCGC-3' and	32
5-CCACCTAAATCGACCTCCGACCG-3';	33
5'-TTATTAGAGGGTGGGGTGGATTGT-3' and	34
5'-CCACCTAAATCAACCTCCAACCA-3';	35
5'-CGCACCCTGCGGCCAGA-3' and	36
5'-AGTGGCCGAGCGGCCGG-3';	37
5'-GCGTTCGTATTTTGCGGTT-3' and	38
5'-CGTACAATAACCGAACGACCGA-3';	39
5'-TGTGATGTGTTTGTATTTTGTGGTT-3' and	40
5'-CCATACAATAACCAAACAACCAA-3';	41
5'-TGGAGGATTTTTTTGCGTACGC-3' and	42
5'-GAACCGAACGCCGCGAA-3';	43
5'-GTTGGAGGATTTTTTTGTGTATGT-3' and	44

-continued

	SEQ ID NO:
5'-CCCAAACCAAACACCACAAA-3';	45
5'-TTAGGTTAGAGGGTTATCGCGT-3' and	46
5'-TAACTAAAAATTCACCTACCGAC-3'; and	47
5'-TAATTTTAGGTTAGAGGGTTATTGT-3' and	48
5'-CACAACCAATCAACAACACA-3'.	49

*Also included are modifications of the above sequences, including SEQ ID NO:26 having the sequence TCAC at the 5' end; SEQ ID NO:27 having the sequence CC added at the 5' end; SEQ ID NO:28 having the sequence 5'-TTATTAGAGGGTGGGGCGGATCGC-3'; SEQ ID NO:29 having the sequence 5'-GACCCCGAACCGCGACCGTAA-3'; SEQ ID NO:30 having the sequence TGG added at the 5' end; and SEQ ID NO:31 having the sequence TACC added at the 5' end. All of these modified primers anneal at 65° C.

Typically, the CpG-containing nucleic acid is in the region of the promoter of a structural gene. For example, the promoter region of tumor suppressor genes have been identified as containing methylated CpG island. The promoter region of tumor suppressor genes, including p16, p15, VHL and E-cadherin, are typically the sequence amplified by PCR in the method of the invention.

Detection and identification of methylated CpG-containing nucleic acid in the specimen may be indicative of a cell proliferative disorder or neoplasia. Such disorders include but are not limited to low grade astrocytoma, anaplastic astrocytoma, glioblastoma, medulloblastoma, colon cancer, lung cancer, renal cancer, leukemia, breast cancer, prostate cancer, endometrial cancer and neuroblastoma. Identification of methylated CpG status is also useful for detection and diagnosis of genomic imprinting, fragile X syndrome and X-chromosome inactivation.

The method of the invention now provides the basis for a kit useful for the detection of a methylated CpG-containing nucleic acid. The kit includes a carrier means being compartmentalized to receive in close confinement therein one or more containers. For example, a first container contains a reagent which modifies unmethylated cytosine, such as sodium bisulfite. A second container contains primers for amplification of the CpG-containing nucleic acid, for example, primers listed above for p16, p15, VHL or E-cadherin.

The above disclosure generally describes the present invention. A more complete understanding can be obtained by reference to the following specific examples which are provided herein for purposes of illustration only and are not intended to limit the scope of the invention.

EXAMPLE 1

DNA and Cell Lines. Genomic DNA was obtained from cell lines, primary tumors and normal tissue as described (Merlo, et al., *Nature Medicine*, 1:686, 1995; Herman, et al., *Cancer Research*, 56:722, 1996; Graff, et al., *Cancer Research*, 55:5195, 1995). The renal carcinoma cell line was kindly provided by Dr. Michael Lehrman of the National Cancer Institute, Bethesda, MD.

Bisulfite Modification. 1 μg of DNA in a volume of 50 μl was denatured by NaOH (final 0.2M) for 10 minutes at 37° C. For samples with nanogram quantities of human DNA, 1 μg of salmon sperm DNA (Sigma) was added as carrier prior to modification. 30 μL of 10 mM hydroquinone (Sigma) and 520 μL of 3M sodium bisulfite (Sigma) pH5, both freshly prepared, were added, mixed, and samples were incubated under mineral oil at 50° C. for 16 hours. Modified DNA was purified using the Wizard™ DNA purification resin according to the manufacturer (Promega), and eluted into 50 μL of water. Modification was completed by NaOH (final 0.3M) treatment for 5 minutes at room temperature, followed by ethanol precipitation.

Genomic Sequencing. Genomic sequencing of bisulfite modified DNA was accomplished using the solid-phase DNA sequencing approach (Myohanen, et al., *DNA Seq.*, 5:1, 1994). 100 ng of bisulfite modified DNA was amplified with p16 gene specific primer 5'-TTTTTAGAGGATTTGAGGGATAGG-3' (sense) (SEQ ID NO:49) and 5'-CTACCTAATTCCAATTCCCCTACA-3' (anti-sense) (SEQ ID NO:50). PCR conditions were as follows: 96° C. for 3 minutes, 80° C. for 3 minutes, 1 U of Taq polymerase (BRL) was added, followed by 35 cycles of 96° C. for 20 seconds, 56° C. for 20 seconds, 72° C. for 90 seconds, followed by 5 minutes at 72° C. The PCR mixture contained 1X buffer (BRL) with 1.5 mM $MgCl_2$, 20 pmols of each primer and 0.2 mM dNTPs. To obtain products for sequencing, a second round of PCR was performed with 5 pmols of nested primers. In this reaction, the sense primer, 5'-GTTTTCCCAGTCACGACAGTATTAGGAGG AAG AAAGAGGAG-3' (SEQ ID NO:51), contains M13–40 sequence (underlined) introduced as a site to initiate sequencing, and the anti-sense primer 5'-TCCAATTCCCCTACAAACTTC-3" (SEQ ID NO:52) is biotinylated to facilitate purification of the product prior to sequencing. PCR was performed as above, for 32 cycles with 2.5 mM $MgCl_2$. All primers for genomic sequencing were designed to avoid any CpGs in the sequence. Biotinylated PCR products were purified using streptavidin coated magnetic beads (Dynal AB, Norway), and sequencing reactions performed with Sequenase™ and M13–40 sequencing primer under conditions specified by the manufacturer (USB).

PCR Amplification. Primer pairs described in Table 1 were purchased from Life Technologies. The PCR mixture contained 1X PCR buffer (16.6 mM ammonium sulfate, 67 mM TRIS pH 8.8, 6.7 mM $MgCl_2$, and 10 mM β-mercaptoethanol), dNTPs (each at 1.25 mM), primers (300 ng/reaction each), and bisulfite-modified DNA (~50 ng) or unmodified DNA (50–100 ng) in a final volume of 50 μL. PCR specific for unmodified DNA also included 5% dimethylsulfoxide. Reactions were hot started at 95° C. for 5 minutes prior to the addition of 1.25 units of Taq polymerase (BRL). Amplification was carried out on a Hybaid OmniGene temperature cycler for 35 cycles (30 seconds at 95° C., 30 seconds at the annealing temperature listed in Table 1, and 30 seconds at 72° C.), followed by a final 4 minute extension at 72° C. Controls without DNA were performed for each set of PCR reactions. 10 μL of each PCR reaction was directly loaded onto non-denaturing 6–8% polyacrylamide gels, stained with ethidium bromide, and directly visualized under UV illumination.

Restriction Analysis. 10 μL of the 50 μL PCR reaction was digested with 10 units of BstUI (New England Biolabs) for 4 hours according to conditions specified by the manufacturer. Restriction digests were ethanol precipitated prior to gel analysis.

EXAMPLE 2

An initial study was required to validate the strategy for MSP for providing assessment of the methylation status of CpG islands. The p16 tumor suppressor (Merlo, et al., supra; Herman, et al., *Cancer Research*, 55:4525, 1995; Gonzalez-Zulueta, et al., *Cancer Res.*, 55:4531, 1995,27) which has been documented to have hypermethylation of a 5' CpG island is associated with complete loss of gene expression in many cancer types, was used as an exemplary gene to determine whether the density of methylation, in key regions to be tested, was great enough to facilitate the primer design disclosed herein. Other than for CpG sites located in recognition sequences for methylation-sensitive enzymes, the density of methylation and its correlation to transcriptional silencing had not yet been established. The genomic sequencing technique was therefore employed to explore this relationship.

FIG. 1 shows genomic sequencing of p16. The sequence shown has the most 5' region at the bottom of the gel, beginning at +175 in relation to a major transcriptional start site (Hara, et al., *Mol. Cell Biol.*, 16:859, 1996). All cytosines in the unmethylated cell line H249 have been converted to thymidine, while all C's in CpG dinucleotides in the methylated cell H157 remains as C, indicating methylation.] enclosed a BstUI site which is at −59 in relation to the transnational start site in Genbank sequence U12818 (Hussussian, et al., *Nat. Genet.*, 8:15, 1994), but which is incorrectly identified as CGCA in sequence X94154 (Hara, et al., supra). This CGCG site represents the 3' location of the sense primer used for p16 MSP.

As has been found for other CpG islands examined in this manner (Myohanen, et al., supra; Park, et al., *Mol. Cell Biol.*, 14:7975, 1994; Reeben, et al., *Gene*, 157:325, 1995), the CpG island of p16 was completely unmethylated in those cell lines and normal tissues previously found to be unmethylated by Southern analysis (FIG. 1)(Merlo, et al., supra; Herman, et al, supra). However, it was extensively methylated in cancer cell lines shown to be methylated by Southern analysis (FIG. 1). In fact, all cytosines within CpG dinucloetides in this region were completely methylated in the cancers lacking p16 transcription. This marked difference in sequence following bisulfite treatment suggested that the method of the invention for specific amplification of either methylated or unmethylated alleles was usefull for identification of methylation patterns in a DNA sample.

Primers were designed to discriminate between methylated and unmethylated alleles following bisulfite treatment, and to discriminate between DNA modified by bisulfite and that which had not been modified. To accomplish this, primer sequences were chosen for regions containing fre-

quent cytosines (to distinguish unmodified from modified DNA), and CpG pairs near the 3' end of the primers (to provide maximal discrimination in the PCR reaction between methylated and unmethylated DNA). Since the two strands of DNA are no longer complementary after bisulfite treatment, primers can be designed for either modified strand. For convenience, primers were designed for the sense strand. The fragment of DNA to be amplified was intentionally small, to allow the assessment of methylation patterns in a limited region and to facilitate the application of this technique to samples, such as paraffin blocks, where amplification of larger fragments is not possible. In Table 1, primer sequences are shown for all genes tested, emphasizing the differences in sequence between the three types of DNA which are exploited for the specificity of MSP. The multiple mismatches in these primers which are specific for these different types of DNA suggest that each primer set should provide amplification only from the intended template.

The primers designed for p16 were tested with DNA from cancer cell lines and normal tissues for which the methylation status had previously been defined by Southern analysis (Merlo, et al., supra; Herman, et al., supra).

FIG. 2, panels A–D, show polyacrylamide gels with the Methylation Specific PCR products of p16. Primer sets used for amplification are designated as unmethylated (U), methylated (M), or unmodified/wild-type (W).* designates the molecular weight marker pBR322-MspI digest. Panel A shows amplification of bisulfite-treated DNA from cancer cell lines and normal lymphocytes, and untreated DNA (from cell line H249). Panel B shows mixing of various amount of H157 DNA with 1 μg of H249 DNA prior to bisulfite treatment to assess the detection sensitivity of MSP for methylated alleles. Modified DNA from a primary lung cancer sample and normal lung are also shown. Panel C shows amplification with the p16-U2 (U) primers, and p16-M2 (M) described in Table 1. Panel D shows the amplified p16 products of panel C restricted with BstUI(+) or not restricted (−). In all cases, the primer set used confirmed the methylation status determined by Southern analysis. For example, lung cancer cell lines U1752 and H157, as well other cell lines methylated at p16, amplified only with the methylated primers (FIG. 2, panel A). DNA from normal tissues (lymphocytes, lung, kidney, breast, and colon) and the unmethylated lung cancer cell lines H209 and H249, amplified only with unmethylated primers (examples in FIG. 2, panel A). PCR with these primers could be performed with or without 5% DMSO. DNA not treated with bisulfite (unmodified) failed to amplify with either set of methylated or unmethylated specific primers, but readily amplified with primers specific for the sequence prior to modification (FIG. 2, panel A). DNA from the cell line H157 after bisulfite treatment also produced a weaker amplification with unmodified primers, suggesting an incomplete bisulfite reaction. However, this unmodified DNA, unlike partially restricted DNA in previous PCR assays relying on methylation sensitive restriction enzymes, is not recognized by the primers specific for methylated DNA. It therefore does not provide a false positive result or interfere with the ability to distinguish methylated from unmethylated alleles.

The sensitivity of MSP for detection of methylated p16 alleles was assessed. DNA from methylated cell lines was mixed with unmethylated DNA prior to bisulfite treatment. 0.1% of riethylated DNA (approximately 50 pg) was consistently detected in an otherwise unmethylated sample (FIG. 2, panel B). The sensitivity limit for the amount of input DNA was determined to be as little as 1 ng of human DNA, mixed with salmon sperm DNA as a carrier detectable by MSP.

Fresh human tumor samples often contain normal and tumor tissue, making the detection of changes specific for the tumor difficult. However, the sensitivity of MSP suggests it would be useful for primary tumors as well, allowing for detection of aberrantly methylated alleles even if they contribute relatively little to the overall DNA in a sample. In each case, while normal tissues were completely unmethylated, tumors determined to be methylated at p16 by Southern analysis also contained methylated DNA detected by MSP, in addition to some unmethylated alleles (examples in FIG. 2, panel B). DNA from paraffin-embedded tumors was also used, and allowed the detection of methylated and unmethylated alleles in these samples (FIG. 2, panel B). To confirm that these results were not unique to this primer set, a second downstream primer for p16 was used which would amplify a slightly larger fragment (Table 1). This second set of primers reproduced the results described above (FIG. 2, panel C), confirming the methylation status defined by Southern blot analysis.

To further verify the specificity of the primers for the methylated alleles and to check specific cytosines for methylation within the region amplified, the differences in sequence between methylated/modified DNA and unmethylated/modified DNA were utilized. Specifically, the BstUI recognition site, CGCG, will remain CGCG if both C's are methylated after bisulfite treatment and amplification, but will become TGTG if unmethylated. Digestion of the amplified products with BstUI distinguishes these two products. Restriction of p16 amplified products illustrates this. Only unmodified products and methylated/modified products, both of which retain the CGCG site, were cleaved by BstUI, while products amplified with unmethylated/modified primers failed to be cleaved (FIG. 2, panel D).

The primer sets discussed above were designed to discriminate heavily methylated CpG islands from unmethylated alleles. To do this, both the upper (sense) and lower (antisense) primers contained CpG sites which could produce methylation-dependent sequence differences after bisulfite treatment. MSP might be employed to examine more regional aspects of CpG island methylation. To examine this, methylation-dependent differences in the sequence of just one primer was tested to determine whether it would still allow discrimination between unmethylated and methylatedp p16 alleles. The antisense primer used for genomic sequencing, 5'-CTACCTAATTCCAATTCCCCTACA-3' (SEQ ID NO:53), was also used as the antisense primer, since the region recognized by the primer contains no CpG sites, and was paired with either a methylated or unmethylated sense primer (Table 1). Amplification of the 313 bp PCR product only occurred with the unmethylated sense primer in H209 and H249 (unmethylated by Southern) and the methylated sense primer in H157 and U1752 (methylated by Southern), indicating that methylation of CpG sites within a defined region can be recognized by specific primers and distinguish between methylated and unmethylated alleles (FIG. 2, panel E). Panel E shows results of testing for regional methylation of CpG islands with MSP, using sense primers p16-U2 (U) and p16-M2 (M), which are methylation specific, and an antisense primer which is not methylation specific.

EXAMPLE 3

The above experiments with p16 were extended to include 3 other genes transcriptionally silenced in human cancers by aberrant hypermethylation of 5' CpG islands.

FIG. 3, panels A–E, show polyacrylamide gels of MSP products from analysis of several genes. Primer sets used for

amplification are not designated as unmethylated (U), methylated (M), or unmodified/wild-type (W). * designates the molecular weight marker pBR322-MspI digest and ** designates the 123 bp molecular weight marker. All DNA samples were bisulfite treated except those designated untreated. Panel A shows the results from MSP for p15. Panel B shows the p15 products restricted with BstUI (+) or not restricted (−). Panel C shows the products of MSP for VHL. Panel D shows the VHL products restricted with BstUI(+) or not restricted (−). Panel E shows the products of MSP for *E-cadherin*.

The cyclin-dependent kinase inhibitor p15 is aberrantly methylated in many leukemic cell lines and primary leukemias (Herman, et al., supra). For p15, MSP again verified the methylation status determined by Southern analysis. Thus, normal lymphocytes and cancer cell lines SW48 and U1752, all unmethylated by Southern analysis (Herman, et al, supra), only amplified with the unmethylated set of primers, while the lung cancer cell line H1618 and leukemia cell line KG1A amplified only with the methylated set of primers (FIG. 3, panel A), consistent with previous Southern analysis results (Herman, et al., supra). The cell line Raji produced a strong PCR product with methylated primers and a weaker band with unmethylated primers. This was the same result for methylation obtained previously by Southern analysis (Herman, et al., supra). Non-cultured leukemia samples, like the primary tumors studied for p16, had amplification with the methylated primer set as well as the unmethylated set. This heterogeneity also matched Southern analysis (Herman, et al., supra). Again, as for p16, differential modification of BstUI restriction sites in the amplified product of p15 was used to verify the specific amplification by MSP (FIG. 3, panel B). Amplified products using methylated primer sets from cell lines H1618 and Raji or unmodified primer sets, were completely cleaved by BstUI, while unmethylated amplified products did not cleave. Primary AML samples, which again only demonstrated cleavage in the methylated product, had less complete cleavage. This suggests a heterogeneity in methylation, arising because in some alleles, many CpG sites within the primer sequences area are methylated enough to allow the methylation specific primers to amplify this region, while other CpG sites are not completely methylated.

Aberrant CpG island promoter region methylation is associated with inactivation of the VHL tumor suppressor gene in approximately 20% of clear renal carcinomas (Herman, et al., *Proc. Natl. Acad. Sci. USA*, 91:9700, 1994). This event, like mutations for VHL (Gnarra, et al., *Nature Genetics*, 7:85, 1994), is restricted to clear renal cancers (Herman, et al, supra). Primers designed for the VHL sequence were used to study DNA from the renal cell cancer line RFX393 which is methylated at VHL by Southern analysis, and the lung cancer cell line U1752 which is unmethylated at this locus (Herman, et al., supra). In each case, the methylation status of VHL determined by MSP confirmed that found by Southern analysis (FIG. 3, panel C), and BstUI restriction site analysis validated the PCR product specificity (FIG. 3, panel D).

The expression of the invasion/metastasis suppressor gene, E-cadherin, is often silenced by aberrant methylation of the 5' promoter in breast, prostate, and many other carcinomas (Graff, et al., supra; Yoshira, et al., *Proc. Natl. Acad. Sci. USA*, 92:7416, 1995). Primers were designed for the *E-cadherin* promoter region to test the use of MSP for this gene. In each case, MSP analysis paralleled Southern blot analysis for the methylation status of the gene (Graff, et al., supra). The breast cancer cell lines MDA-MB-231, HS578t, and the prostate cancer cell lines DuPro and TSUPrI, all heavily methylated by Southern, displayed prominent methylation. MCF7, T47D, PC-3, and LNCaP, all unmethylated by Southern, showed no evidence for methylation in the sensitive MSP assay (FIG. 3, panel E). MSP analysis revealed the presence of unmethylated alleles in Hs578t, TSUPrI and DuPro consistent with a low percentage of unmethylated alleles in these cell lines previously detected by Southern analysis (Graff, el al., supra). BstUI restriction analysis again confirmed the specificity of the PCR amplification.

TABLE 1

PCR primers used for Methylation Specific PCR

Primer Set	Sense primer* (5'–3')	Antisense primer* (5'–3')	Size (bp)	Anneal temp.	Genomic Position†
p16-W†	CAGAGGGTGGGGCGACCGC	CGGGCCGCGGCCGTGG	140	65° C.	+171
p16-M	TTATTAGAGGGTGGGGCGGATCGC	GACCCCGAACCGCGACCGTAA	150	65° C.	+167
p16-U	TTATTAGAGGGTGGGGTGGATTGT	CAACCCCAAACCACAACCATAA	151	60° C.	+167
p16-M2	TTATTAGAGGGTGGGGCGGATCGC	CCACCTAAATCGACCTCCGACCG	234	65° C.	+167
p16-U2	TTATTAGAGGGTGGGGTGGATTGT	CCACCTAAATCAACCTCCAACCA	234	60° C.	+167
p15-W	CGCACCCTGCGGCCAGA	AGTGGCCGAGCGGCCGG	137	65° C.	+46
p15-M	GCGTTCGTATTTTGCGGTT	CGTACAATAACCGAACGACCGA	148	60° C.	+40
p15-U	TGTGATGTGTTTGTATTTTGTGGTT	CCATACAATAACCAAACAACCAA	154	60° C.	+34
VHL-M	TGGAGGATTTTTTTGCGTACGC	GAACCGAACGCCGCGAA	158	60° C.	−116
VHL-U	GTTGGAGGATTTTTTTGTGTATGT	CCCAAACCAAACACCACAAA	165	60° C.	−118
Ecad-M	TTAGGTTAGAGGGTTATCGCGT	TAACTAAAAATTCACCTACCGAC	116	57° C.	−205
Ecad-U	TAATTTTAGGTTAGAGGGTTATTGT	CACAACCAATCAACAACACA	97	53° C.	−210

*Sequence differences between modified primers and unmodified DNA are boldface, and differences between methylated/modified acid unmethylated/modified are underlined.

†Primers were placed near the transcriptional strat site. Genomic position is the location of the 5' nucleotide of the sense primer in relation to the major transcriptional start site defined in the following references and Genbank accession numbers: p16(most 3' site) X94154 (E. Hara, et al., Mol. Cell Biol., 16:859, 1996), p15 S75756 (J. Jen, et al., Cancer Res., 54: 6353 1994), VHL U19763 (I. Kuzmin, et al., Oncogene, 10:2185 1995), and E-cadherin 34545 (M. J. Bussemakers, et al., Biochem. Biophys. Res. Commun., 203: 1284 1994).

†W represents unmodified, or wild-type primers, M represents methylated-specific primers, and U represents unmethylated-specific primers. (SEQ ID NO:26–48)

Although the invention has been described with reference to the presently preferred embodiments, it should be understood that various modifications can be made without departing from the spirit of the invention. Accordingly, the invention is limited only by the following claims.

SEQUENCE LISTING

(1) GENERAL INFORMATION:

(i i i) NUMBER OF SEQUENCES: 52

(2) INFORMATION FOR SEQ ID NO:1:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 20 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: DNA

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:1:

GCGGTCCGCC CCACCCTCTG 20

(2) INFORMATION FOR SEQ ID NO:2:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 16 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: DNA

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:2:

CCACGGCCGC GGCCCG 16

(2) INFORMATION FOR SEQ ID NO:3:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 24 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: DNA

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:3:

GCGATCCGCC CCACCCTCTA ATAA 24

(2) INFORMATION FOR SEQ ID NO:4:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 21 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: DNA

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:4:

TTACGGTCGC GGTTCGGGGT C 21

(2) INFORMATION FOR SEQ ID NO:5:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 24 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: DNA

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:5:

-continued

ACAATCCACC CCACCCTCTA ATAA 24

(2) INFORMATION FOR SEQ ID NO:6:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 22 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:6:

TTATGGTTGT GGTTTGGGGT TG 22

(2) INFORMATION FOR SEQ ID NO:7:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 24 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:7:

GCGATCCGCC CCACCCTCTA ATAA 24

(2) INFORMATION FOR SEQ ID NO:8:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 23 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:8:

CGGTCGGAGG TCGATTTAGG TGG 23

(2) INFORMATION FOR SEQ ID NO:9:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 24 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:9:

ACAATCCACC CCACCCTCTA ATAA 24

(2) INFORMATION FOR SEQ ID NO:10:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 23 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:10:

TGGTTGGAGG TTGATTTAGG TGG 23

(2) INFORMATION FOR SEQ ID NO:11:

-continued

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 17 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: DNA

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:11:

TCTGGCCGCA GGGTGCG 17

(2) INFORMATION FOR SEQ ID NO:12:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 17 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: DNA

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:12:

CCGGCCGCTC GGCCACT 17

(2) INFORMATION FOR SEQ ID NO:13:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 19 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: DNA

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:13:

AACCGCAAAA TACGAACGC 19

(2) INFORMATION FOR SEQ ID NO:14:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 22 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: DNA

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:14:

TCGGTCGTTC GGTTATTGTA CG 22

(2) INFORMATION FOR SEQ ID NO:15:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 25 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: DNA

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:15:

AACCACAAAA TACAAACACA TCACA 25

(2) INFORMATION FOR SEQ ID NO:16:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 23 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

-continued

(i i) MOLECULE TYPE: DNA

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:16:

TTGGTTGTTT GGTTATTGTA TGG 23

(2) INFORMATION FOR SEQ ID NO:17:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 22 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: DNA

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:17:

GCGTACGCAA AAAAATCCTC CA 22

(2) INFORMATION FOR SEQ ID NO:18:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 17 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: DNA

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:18:

TTCGCGGCGT TCGGTTC 17

(2) INFORMATION FOR SEQ ID NO:19:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 24 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: DNA

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:19:

ACATACACAA AAAAATCCTC CAAC 24

(2) INFORMATION FOR SEQ ID NO:20:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 20 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: DNA

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:20:

TTTGTGGTGT TTGGTTTGGG 20

(2) INFORMATION FOR SEQ ID NO:21:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 22 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: DNA

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:21:

-continued

ACGCGATAAC CCTCTAACCT AA 22

(2) INFORMATION FOR SEQ ID NO:22:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 23 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:22:

GTCGGTAGGT GAATTTTTAG TTA 23

(2) INFORMATION FOR SEQ ID NO:23:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 25 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:23:

ACAATAACCC TCTAACCTAA AATTA 25

(2) INFORMATION FOR SEQ ID NO:24:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 20 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:24:

TGTGTTGTTG ATTGGTTGTG 20

(2) INFORMATION FOR SEQ ID NO:25:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 20 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:25:

CAGAGGGTGG GGCGGACCGC 20

(2) INFORMATION FOR SEQ ID NO:26:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 16 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:26:

CGGGCCGCGG CCGTGG 16

(2) INFORMATION FOR SEQ ID NO:27:

-continued

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 24 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: DNA

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:27:

TTATTAGAGG GTGGGGCGGA TCGC 24

(2) INFORMATION FOR SEQ ID NO:28:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 21 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: DNA

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:28:

GACCCCGAAC CGCGACCGTA A 21

(2) INFORMATION FOR SEQ ID NO:29:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 24 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: DNA

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:29:

TTATTAGAGG GTGGGGTGGA TTGT 24

(2) INFORMATION FOR SEQ ID NO:30:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 22 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: DNA

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:30:

CAACCCCAAA CCACAACCAT AA 22

(2) INFORMATION FOR SEQ ID NO:31:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 24 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: DNA

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:31:

TTATTAGAGG GTGGGGTGGA TTGT 24

(2) INFORMATION FOR SEQ ID NO:32:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 23 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

-continued

(i i) MOLECULE TYPE: DNA

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:32:

CCACCTAAAT CGACCTCCGA CCG 23

(2) INFORMATION FOR SEQ ID NO:33:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 24 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: DNA

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:33:

TTATTAGAGG GTGGGGTGGA TTGT 24

(2) INFORMATION FOR SEQ ID NO:34:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 23 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: DNA

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:34:

CCACCTAAAT CAACCTCCAA CCA 23

(2) INFORMATION FOR SEQ ID NO:35:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 17 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: DNA

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:35:

CGCACCCTGC GGCCAGA 17

(2) INFORMATION FOR SEQ ID NO:36:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 17 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: DNA

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:36:

AGTGGCCGAG CGGCCGG 17

(2) INFORMATION FOR SEQ ID NO:37:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 19 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: DNA

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:37:

-continued

GCGTTCGTAT TTTGCGGTT 19

(2) INFORMATION FOR SEQ ID NO:38:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 22 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: DNA

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:38:

CGTACAATAA CCGAACGACC GA 22

(2) INFORMATION FOR SEQ ID NO:39:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 25 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: DNA

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:39:

TGTGATGTGT TTGTATTTTG TGGTT 25

(2) INFORMATION FOR SEQ ID NO:40:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 23 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: DNA

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:40:

CCATACAATA ACCAAACAAC CAA 23

(2) INFORMATION FOR SEQ ID NO:41:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 22 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: DNA

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:41:

TGGAGGATTT TTTTGCGTAC GC 22

(2) INFORMATION FOR SEQ ID NO:42:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 17 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: DNA

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:42:

GAACCGAACG CCGCGAA 17

(2) INFORMATION FOR SEQ ID NO:43:

-continued

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 24 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: DNA

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:43:

GTTGGAGGAT TTTTTTGTGT ATGT 24

(2) INFORMATION FOR SEQ ID NO:44:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 20 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: DNA

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:44:

CCCAAACCAA ACACCACAAA 20

(2) INFORMATION FOR SEQ ID NO:45:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 22 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: DNA

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:45:

TTAGGTTAGA GGGTTATCGC GT 22

(2) INFORMATION FOR SEQ ID NO:46:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 23 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: DNA

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:46:

TAACTAAAAA TTCACCTACC GAC 23

(2) INFORMATION FOR SEQ ID NO:47:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 25 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: DNA

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:47:

TAATTTTAGG TTAGAGGGTT ATTGT 25

(2) INFORMATION FOR SEQ ID NO:48:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 20 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

-continued

(i i) MOLECULE TYPE: DNA

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:48:

CACAACCAAT CAACAACACA 20

(2) INFORMATION FOR SEQ ID NO:49:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 24 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: DNA

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:49:

TTTTTAGAGG ATTTGAGGGA TAGG 24

(2) INFORMATION FOR SEQ ID NO:50:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 24 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: DNA

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:50:

CTACCTAATT CCAATTCCCC TACA 24

(2) INFORMATION FOR SEQ ID NO:51:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 41 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: DNA

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:51:

GTTTTCCCAG TCACGACAGT ATTAGGAGGA AGAAAGAGGA G 41

(2) INFORMATION FOR SEQ ID NO:52:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 21 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: DNA

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:52:

TCCAATTCCC CTACAAACTT C 21

What is claimed is:

1. A method for detecting a methylated CpG-containing nucleic acid comprising:

contacting a nucleic acid-containing specimen with an agent that modifies unmethylated cytosine,

amplifying the CpG-containing nucleic acid in the specimen by means of CpG-specific oligonucleotide primers, wherein the oligonucleotide primers distinguish between modified methylated and nonmethylated nucleic acid, and

detecting the methylated nucleic acid based on the presence or absence of amplification products produced in said amplifying step.

2. The method of claim 1, wherein the amplifying step is the polymerase chain reaction (PCR).

3. The method of claim 1, wherein the modifying agent is bisulfite.

4. The method of claim 1, wherein cytosine is modified to uracil.

5. The method of claim 1, wherein the CpG-containing nucleic acid is in a promoter region.

6. The method of claim 5, wherein the promoter is a tumor suppressor gene promoter.

7. The method of claim 6, wherein the tumor suppressor gene is selected from the group consisting of p16, p15, E-cadherin, and VHL.

8. The method of claim 1, wherein the specimen is from a tissue selected from the group consisting of brain, colon, urogenital, lung, renal, hematopoietic, breast, thymus, testis, ovarian, and uterine.

9. The method of claim 1, further comprising contacting the nucleic acid with a methylation sensitive restriction endonuclease.

10. The method of claim 9, wherein the restriction endonuclease is selected from the group consisting ofMspI, HpaII, BssHII, BstUI and NotI.

11. The method of claim 1, wherein the presence of methylated CpG-containing nucleic acid in the specimen is indicative of a cell proliferative disorder.

12. The method of claim 11, wherein the disorder is selected from the group consisting of low grade astrocytoma, anaplastic astrocytoma, glioblastoma, medulloblastoma, colon cancer, lung cancer, renal cancer, leukemia, breast cancer, prostate cancer, endometrial cancer and neuroblastoma.

13. The method of claim 1, wherein the primer hybridizes with a target polynucleotide sequence having the sequence selected from the group consisting of SEQ ID NO:1–23 and SEQ ID NO:24.

14. The method of claim 1, wherein the primers are selected from the group consisting of SEQ ID NO:25–47 and SEQ ID NO:48.

15. A kit useful for the detection of a methylated CpG-containing nucleic acid comprising carrier means being compartmentalized to receive in close confinement therein one or more containers comprising a first container containing a reagent which modifies unmethylated cytosine and a second container containing primers for amplification of the CpG-containing nucleic acid, wherein the primers distinguish between modified methylated and nonmethylated nucleic acid.

16. The kit of claim 15, wherein the modifying reagent is bisulfite.

17. The kit of claim 15, wherein said reagent modifies cytosine to uracil.

18. The kit of claim 15, wherein the primer hybridizes with a target polynucleotide sequence having the sequence selected from the group consisting of SEQ ID NO:1–23 and SEQ ID NO:24.

19. The kit of claim 15, wherein the primers are selected from the group consisting of SEQ ID NO:25–47 and 48.

20. Isolated oligonucleotide primer(s) for detection of a methylated CpG-containing nucleic acid wherein the primer hybridizes with a target polynucleotide sequence having the sequence selected from the group consisting of SEQ ID NO:1–23 and SEQ ID NO:24.

21. The primers of claim 20, wherein the primer pairs are selected from the group consisting of SEQ ID NO:25–47 and 48.

22. A kit for the detection of methylated CpG-containing nucleic acid from a sample comprising:

a) a reagent that modifies unmethylated cytosine nucleotides;

b) a wild-type unmodified control nucleic acid;

c) primers for the amplification of unmethylated CpG-containing nucleic acid;

d) primers for the amplification of methylated CpG-containing nucleic acid; and

e) primers for the amplification of control unmodified nucleic acid,

wherein the primers for the amplification of unmethylated CpG-containing nucleic acid and methylated CpG-containing nucleic acid distinguish between modified methylated and nonmethylated nucleic acid.

23. The kit of claim 22, further comprising nucleic acid amplification buffer.

24. The kit of claim 22, wherein the reagent that modifies unmethylated cytosine is bisulfite.

25. The kit of claim 22, wherein primers hybridize with a target polynucleotide sequence having the sequence selected from the group consisting of SEQ ID NO:1–23 and SEQ ID NO:24.

26. The kit of claim 22, wherein the primers are selected from the group consisting of SEQ ID NO:25–47 and SEQ ID NO:48.

27. A method for detecting a methylated CpG-containing nucleic acid comprising:

contacting a nucleic acid-containing specimen with bisulfite to modify unmethylated cytosine,

amplifying the CpG-containing nucleic acid in the specimen by means of CpG-specific oligonucleotide primers, wherein the oligonucleotide primers distinguish between modified methylated and nonmethylated nucleic acid and

detecting the methylated nucleic acid based on the presence or absence of amplification products produced in said amplifying step.

* * * * *

US 20030232351A1

(19) **United States**

(12) **Patent Application Publication** (10) **Pub. No.: US 2003/0232351 A1**

Feinberg (43) **Pub. Date: Dec. 18, 2003**

(54) **METHODS FOR ANALYZING METHYLATED CPG ISLANDS AND GC RICH REGIONS**

(76) Inventor: **Andrew P. Feinberg**, Lutherville, MD (US)

Correspondence Address:
GRAY CARY WARE & FREIDENRICH LLP
4365 EXECUTIVE DRIVE
SUITE 1100
SAN DIEGO, CA 92121-2133 (US)

(21) Appl. No.: **10/308,862**

(22) Filed: **Dec. 2, 2002**

Related U.S. Application Data

(60) Provisional application No. 60/338,888, filed on Nov. 30, 2001.

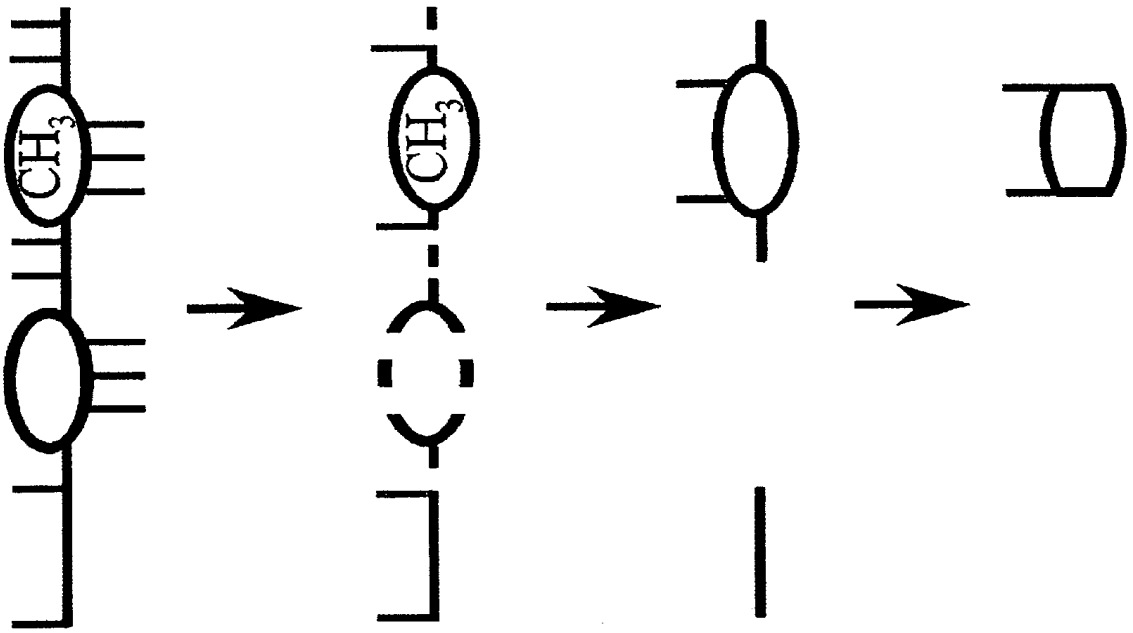
Publication Classification

(51) **Int. Cl.**[7] **C12Q 1/68**

(52) **U.S. Cl.** **435/6**

(57) **ABSTRACT**

The present invention provides CpG islands and GC rich regions and methods for identifying methylation states for these CpG islands and GC rich regions. The present invention also provides methods for identifying genes regulated by these CpG islands and GC rich regions, and provides methods for identifying a population of CpG islands and GC rich regions in a genome.



Step 1
Mse I digestion
Hpa II digestion
Clone larger Mse I fragments

Step 2
Eag I digestion
Subclone Eag I Fragments

Figure 1

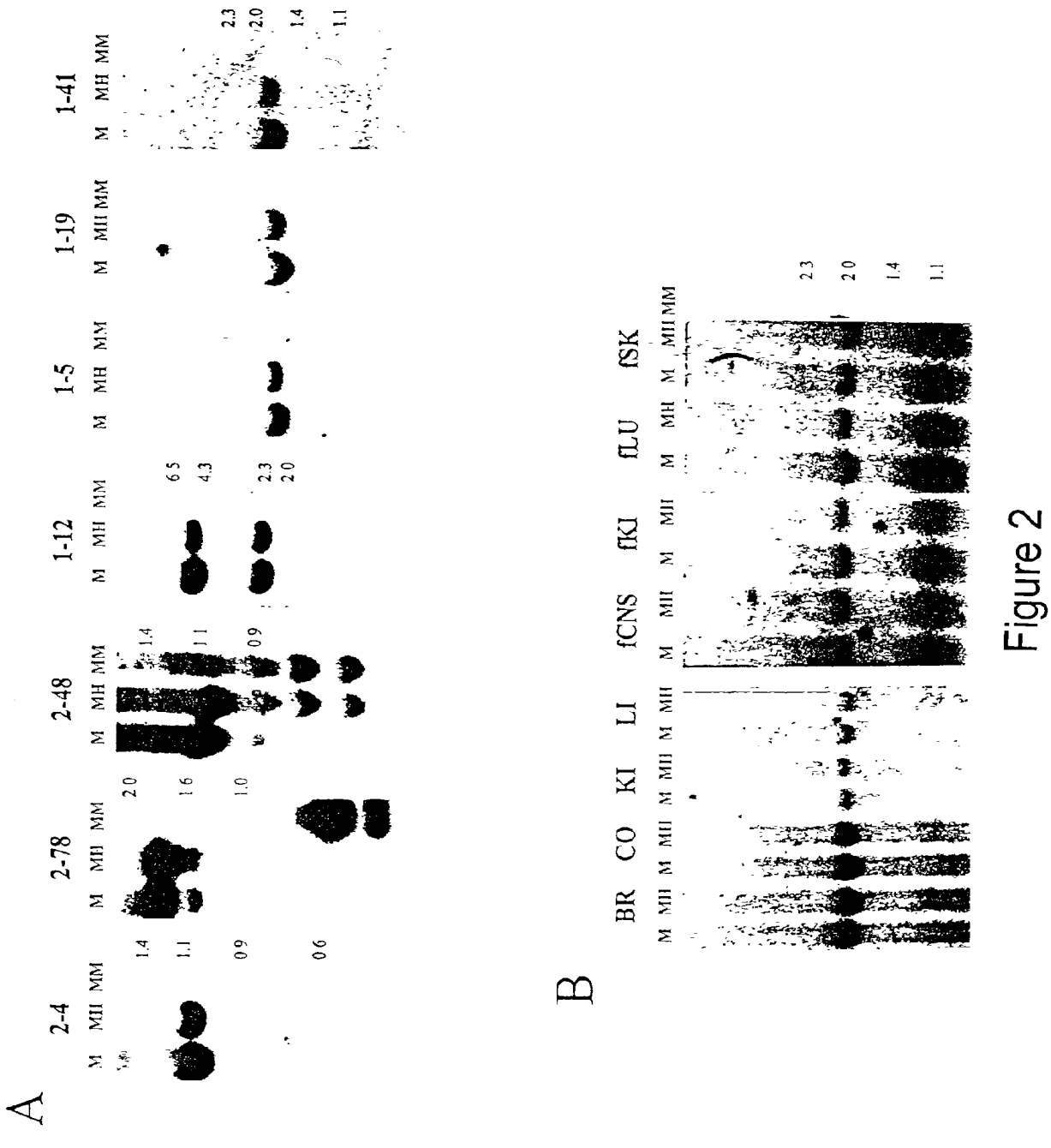

A
2-4
2-78
2-48
1-12
1-5
1-19
1-41
M MII MM
1.4
1.1
0.9
0.6
2.0
1.6
1.0
1.4
1.1
0.9
6.5
4.3
2.3
2.0
2.3
2.0
1.4
1.1
B
BR CO KI LI fCNS fKJ fLU fSK
M MII
2.3
2.0
1.4
1.1

Figure 2

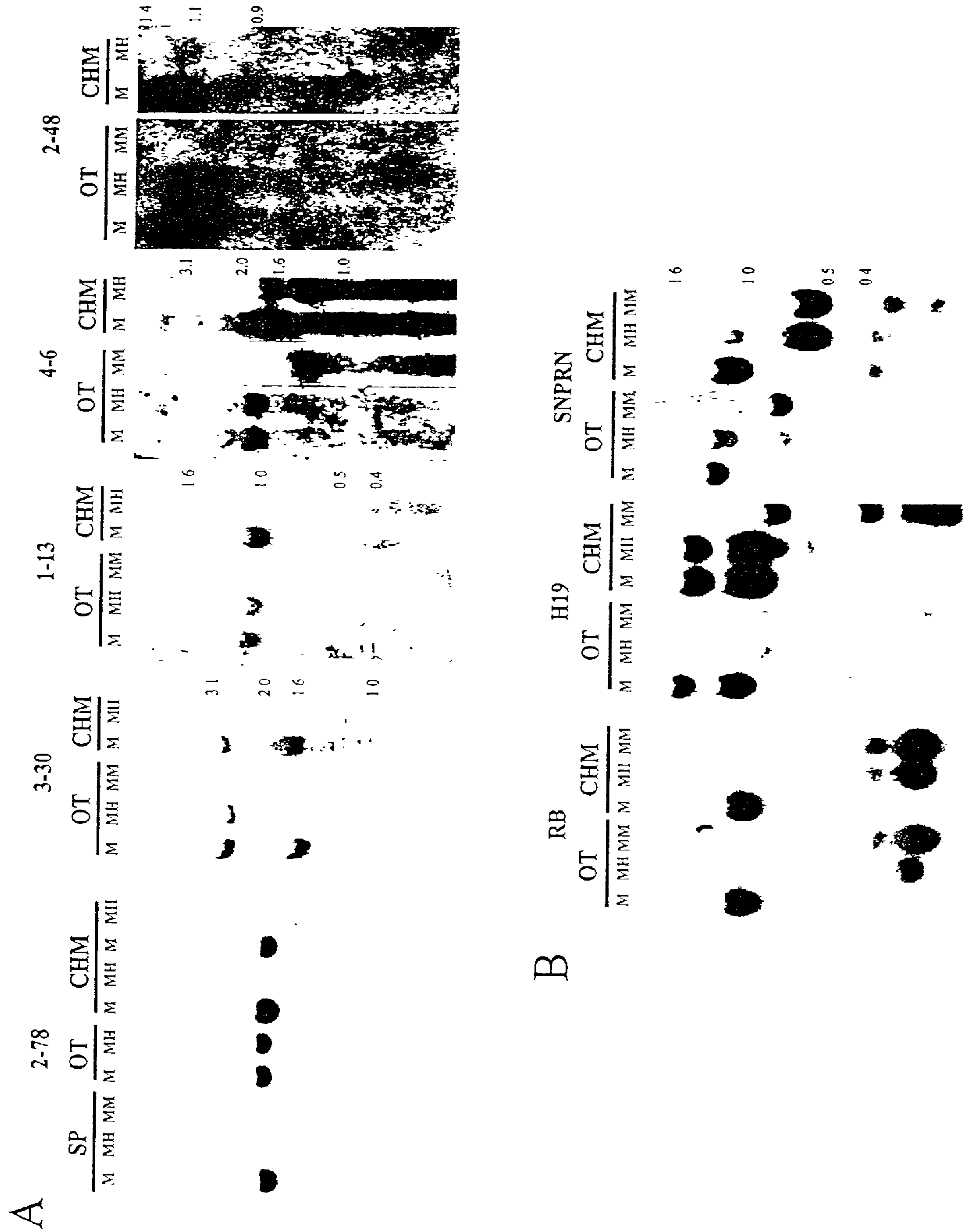


Figure 3

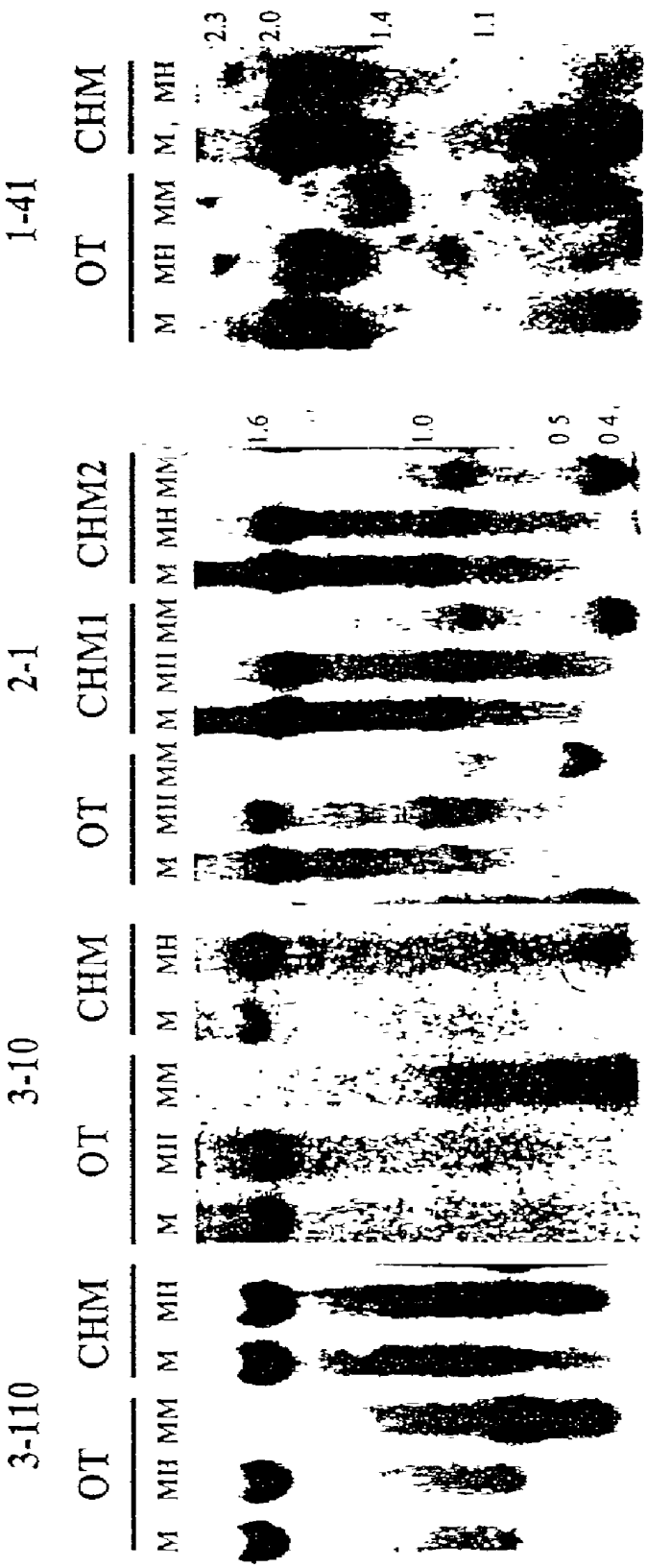

3-110
OT
M MH MM
CHM
M MH
3-10
OT
M MH MM
CHM
M MH
2-1
OT
M MH MM
CHM1
M MH MM
CHM2
M MH MM
1-41
OT
M MH MM
CHM
M, MH
2.3
2.0
1.4
1.1
1.6
1.0
0.5
0.4

Figure 4

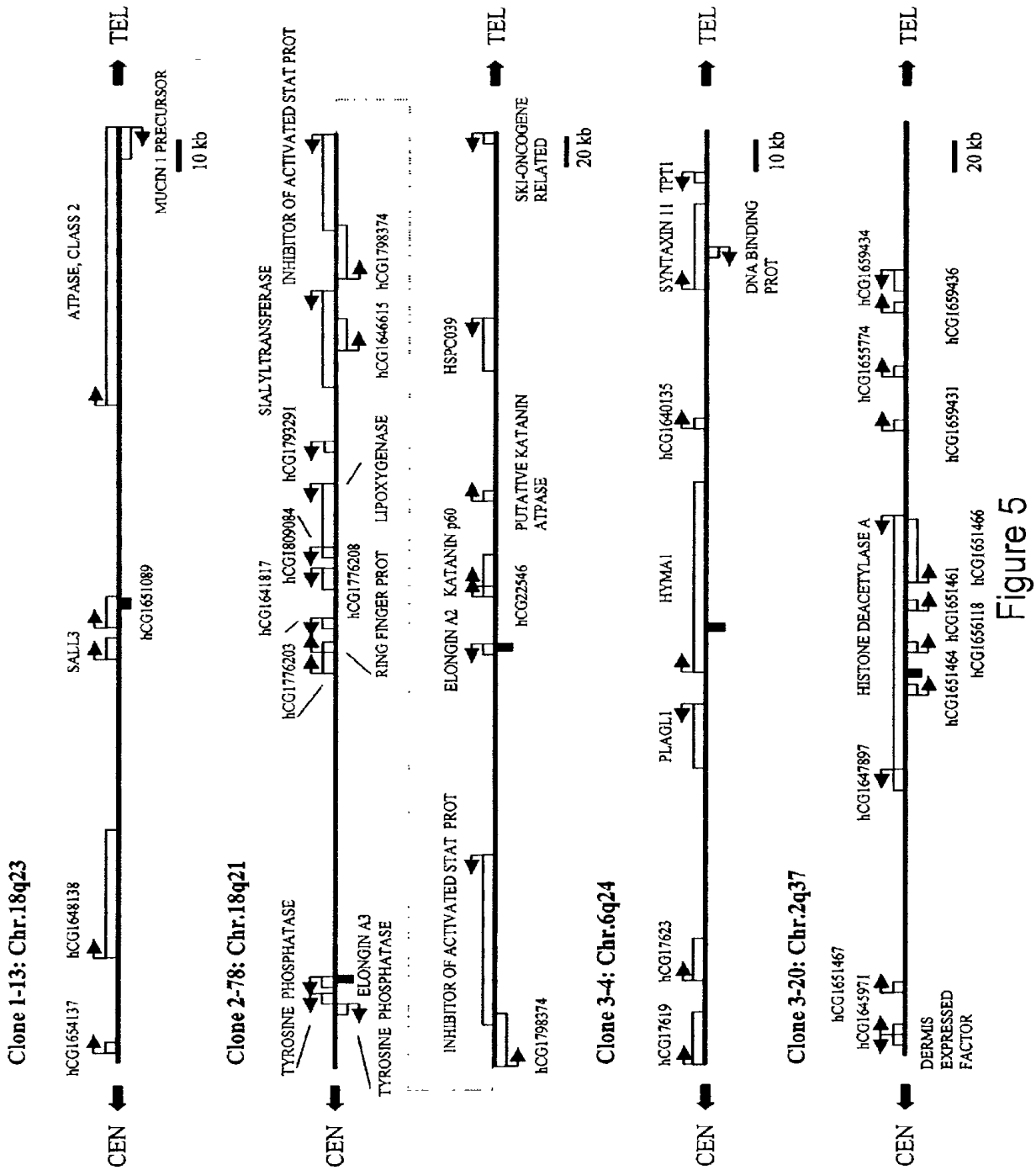

Clone 1-13: Chr.18q23
CEN
TEL
hCG1654137
hCG1648138
SALL3
ATPASE, CLASS 2
MUCIN 1 PRECURSOR
hCG1651089
10 kb
Clone 2-78: Chr.18q21
TYROSINE PHOSPHATASE
ELONGIN A3
TYROSINE PHOSPHATASE
hCG1641817
hCG1776203
hCG1809084
hCG1793291
SIALYLTRANSFERASE
INHIBITOR OF ACTIVATED STAT PROT
hCG1776208
RING FINGER PROT
LIPOXYGENASE
hCG1646615
hCG1798374
INHIBITOR OF ACTIVATED STAT PROT
ELONGIN A2
KATANIN p60
HSPC039
hCG22546
PUTATIVE KATANIN ATPASE
SKI-ONCOGENE RELATED
hCG1798374
20 kb
Clone 3-4: Chr.6q24
hCG17619
hCG17623
PLAGL1
HYMA1
hCG1640135
SYNTAXIN 11
TPT1
DNA BINDING PROT
10 kb
Clone 3-20: Chr.2q37
hCG1651467
hCG1645971
DERMIS EXPRESSED FACTOR
hCG1647897
HISTONE DEACETYLASE A
hCG1651464
hCG1651461
hCG1656118
hCG1651466
hCG1655774
hCG1659434
hCG1659431
hCG1659436
20 kb

Figure 5

```
   1 atggcggcagggtccactacgctgcgcgcagtggggaagctgcag
     M  A  A  G  S  T  T  L  R  A  V  G  K  L  Q
  46 gtgcgtctggccactaagacggagccgaaaaagctagagaaatat
     V  R  L  A  T  K  T  E  P  K  K  L  E  K  Y
  91 ttgcagaaactctccgccttgcccatgaccgcagacatcctggcg
     L  Q  K  L  S  A  L  P  M  T  A  D  I  L  A
 136 gagactggaatcagaaagacggtgaagcgcctgcggaagcaccag
     E  T  G  I  R  K  T  V  K  R  L  R  K  H  Q
 181 cacgtgggcgactttgccagagacttagcggcccggtggaagaag
     H  V  G  D  F  A  R  D  L  A  A  R  W  K  K
 226 ctggtgctcgtggaccgaaacaccgggcctgacccgcaggaccct
     L  V  L  V  D  R  N  T  G  P  D  P  Q  D  P
 271 gaggagagcgcttcccgacagcgcttcggggaggctcttcaggag
     E  E  S  A  S  R  Q  R  F  G  E  A  L  Q  E
 316 cgggaaaaggcctggggcttcccagaaaacgcgacggccccccagg
     R  E  K  A  W  G  F  P  E  N  A  T  A  P  R
 361 agcccatctcacagccctgagcacagacggacagcacgcagaaca
     S  P  S  H  S  P  E  H  R  R  T  A  R  R  T
 406 cctccggggcaacagagacctcacccgaggtctcccagtcgcgag
     P  P  G  Q  Q  R  P  H  P  R  S  P  S  R  E
 451 cccagagccgagagaaagcgccccagaatggccccagctgattcc
     P  R  A  E  R  K  R  P  R  M  A  P  A  D  S
 496 ggcccccatcgggaccctccaacgcgcaccgctcccctcccgatg
     G  P  H  R  D  P  P  T  R  T  A  P  L  P  M
 541 cccgagggccctgagcccgctgtgcccggggagcaacccggaaga
     P  E  G  P  E  P  A  V  P  G  E  Q  P  G  R
 586 ggccacgctcacgccgctcagggcgggcctctgctgggtcaaggc
     G  H  A  H  A  A  Q  G  G  P  L  L  G  Q  G
 631 tgccagggccaacccccaggggggaagcggtggggagccacagcaag
     C  Q  G  Q  P  Q  G  E  A  V  G  S  H  S  K
 676 gggcacaaatcgtcccgcggggcttcggctcagaaatcgcctcct
     G  H  K  S  S  R  G  A  S  A  Q  K  S  P  P
 721 gtccaggaaagccagtcagagaggctgcaggcggccggcgctgat
     V  Q  E  S  Q  S  E  R  L  Q  A  A  G  A  D
 766 tccgccgggccgaaaacggtgcccagccatgtcttctcggagctc
     S  A  G  P  K  T  V  P  S  H  V  F  S  E  L
 811 tgggacccctcagaggcctggatgcaggccaactacgatctgctg
     W  D  P  S  E  A  W  M  Q  A  N  Y  D  L  L
 856 tccgcttttgaggccatgacctcccaggcaaacccagaagcactc
     S  A  F  E  A  M  T  S  Q  A  N  P  E  A  L
 901 tccgcgccagcgctccaggaggaagctgctttccctggacgcaga
     S  A  P  A  L  Q  E  E  A  A  F  P  G  R  R
 946 gtgaacgctaagatgccggtgtactcgggctccaggcctgcctgc
     V  N  A  K  M  P  V  Y  S  G  S  R  P  A  C
 991 cagctccaggtgccgacgctgcgccagcagtgcctccgggtgcct
     Q  L  Q  V  P  T  L  R  Q  Q  C  L  R  V  P
1036 aggaacaatccggacgccctcggcgacgtggaaggggtcccctac
     R  N  N  P  D  A  L  G  D  V  E  G  V  P  Y
1081 tcggttcttgaacccgttctggaagggtggacgcccgatcagctg
     S  V  L  E  P  V  L  E  G  W  T  P  D  Q  L
1126 taccgcacagagaaagacaatgccgcactcgctcgagagacagat
     Y  R  T  E  K  D  N  A  A  L  A  R  E  T  D
1171 gaattatggaggattcattgcctccaggacttcaaggaagaaaag
     E  L  W  R  I  H  C  L  Q  D  F  K  E  E  K
1216 ccacaggagcacgagtcttggcgggagctgtacctgcggcttcgg
     P  Q  E  H  E  S  W  R  E  L  Y  L  R  L  R
1261 gacgcccgagagcagcggctgcgagtagtgaccacgaaaatccga
     D  A  R  E  Q  R  L  R  V  V  T  T  K  I  R
1306 tccgcacgtgaaaacaaacccagcggccgacagacaaagatgatc
     S  A  R  E  N  K  P  S  G  R  Q  T  K  M  I
1351 tgtttcaactctgtggccaagacgccttatgatgcttccaggagg
     C  F  N  S  V  A  K  T  P  Y  D  A  S  R  R
1396 caagagaagtctgcaggagccgctgaccccggaaatggagagatg
     Q  E  K  S  A  G  A  A  D  P  G  N  G  E  M
1441 gagccagccccccaagcccgcaggaagcagccaggctccctccggc
     E  P  A  P  K  P  A  G  S  S  Q  A  P  S  G
1486 ctcggggacggcgacggcggcagcgtgagcggcggcggcagcagc
     L  G  D  G  D  G  G  S  V  S  G  G  G  S  S
1531 aaccggcacgcggcgcccgcggacaaaacccgaaaacaggctgcc
     N  R  H  A  A  P  A  D  K  T  R  K  Q  A  A
1576 aagaaagtggccccgctgatggccaaggcaattcgagactacaag
     K  K  V  A  P  L  M  A  K  A  I  R  D  Y  K
1621 ggaagattctcccgacgataa 1884
     G  R  F  S  R  R  *
```

Figure 6

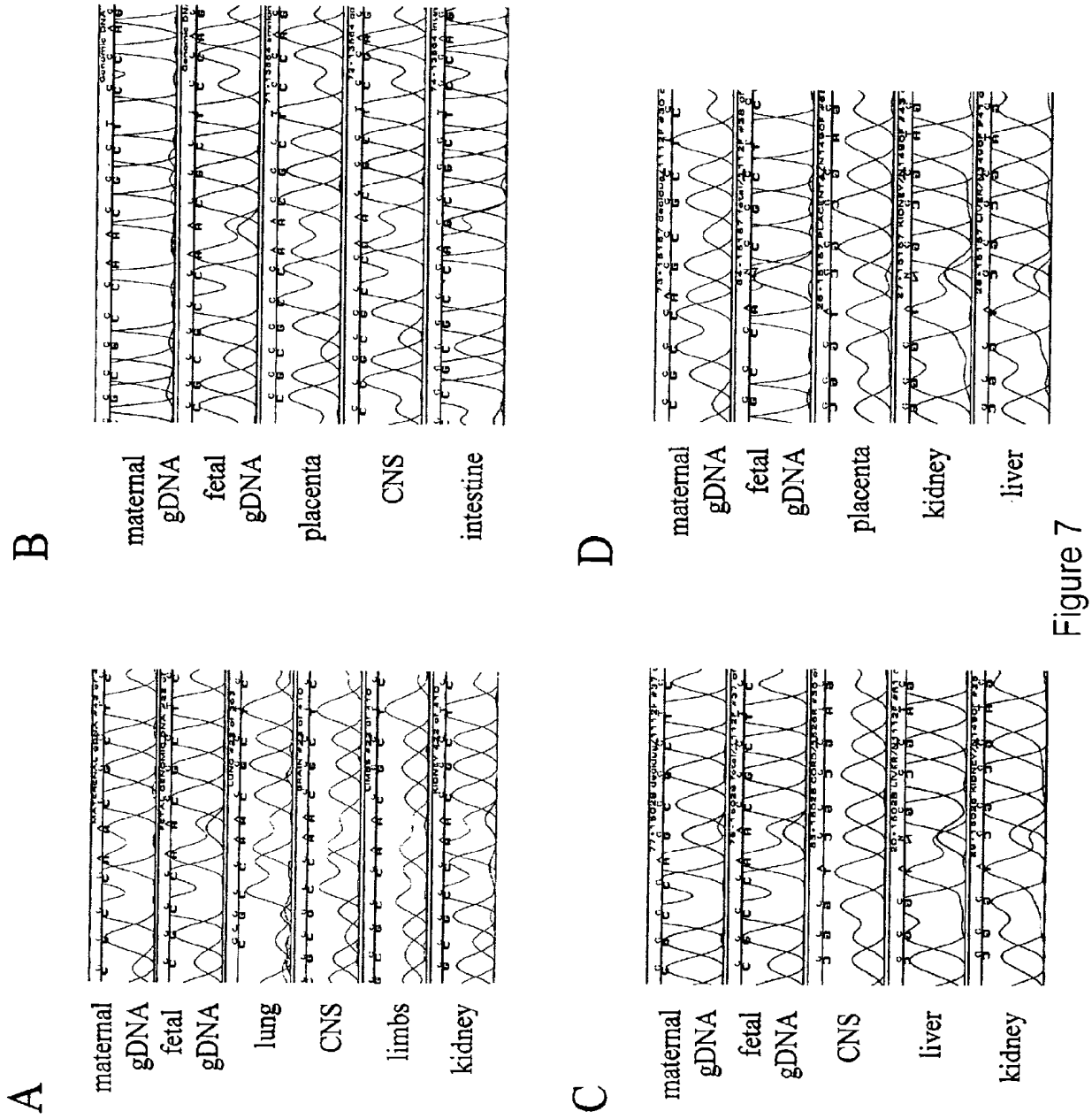

A
maternal gDNA
fetal gDNA
lung
CNS
limbs
kidney
B
maternal gDNA
fetal gDNA
placenta
CNS
intestine
C
maternal gDNA
fetal gDNA
CNS
liver
kidney
D
maternal gDNA
fetal gDNA
placenta
kidney
liver

Figure 7

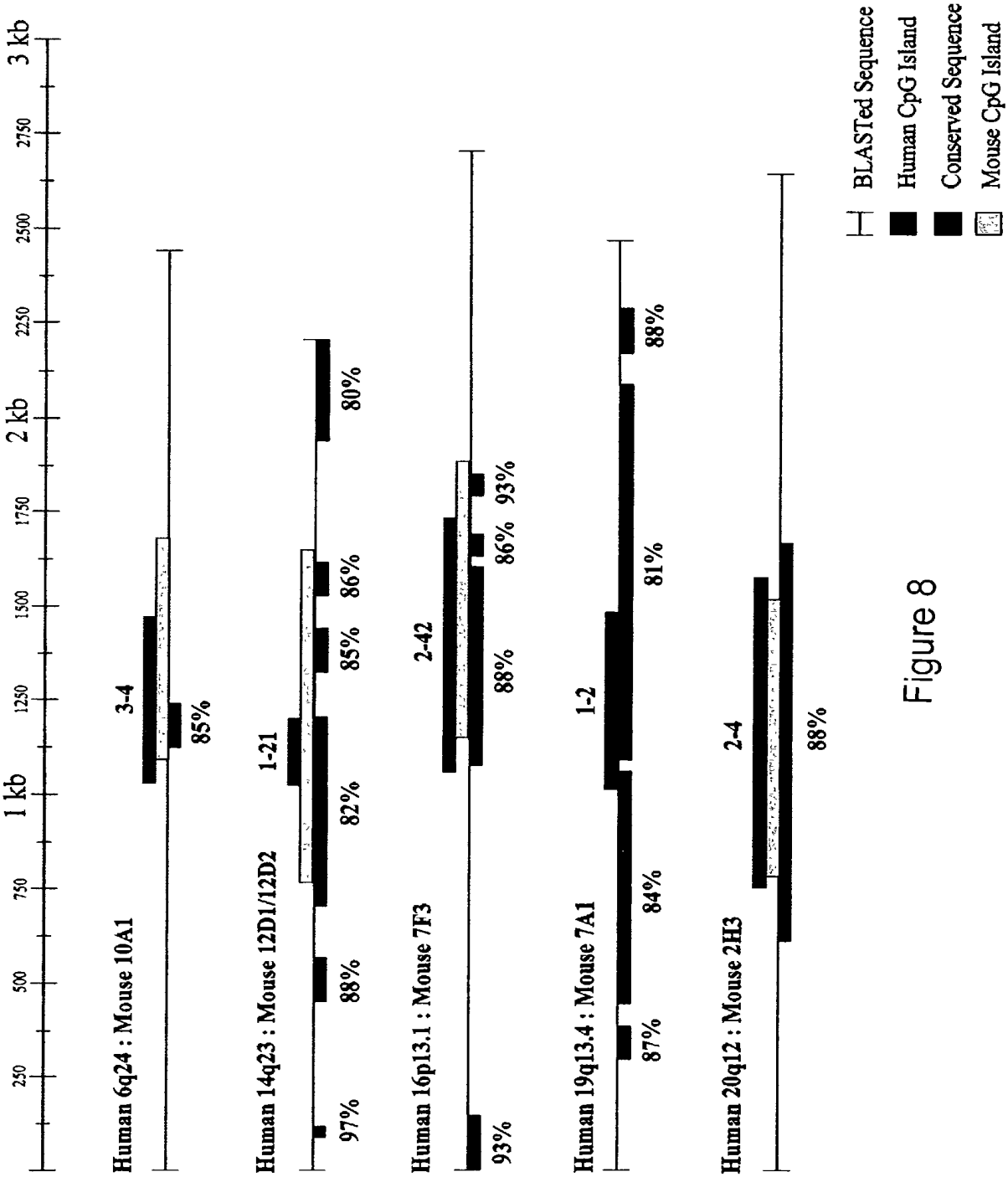

250
500
750
1 kb
1250
1500
1750
2 kb
2250
2500
2750
3 kb
BLASTed Sequence
Human CpG Island
Conserved Sequence
Mouse CpG Island
Human 6q24 : Mouse 10A1
3-4
85%
Human 14q23 : Mouse 12D1/12D2
1-21
97%
88%
82%
85%
86%
80%
Human 16p13.1 : Mouse 7F3
2-42
93%
88%
86%
93%
Human 19q13.4 : Mouse 7A1
1-2
87%
84%
81%
88%
Human 20q12 : Mouse 2H3
2-4
88%

Figure 8

METHODS FOR ANALYZING METHYLATED CPG ISLANDS AND GC RICH REGIONS

RELATED APPLICATION DATA

[0001] This application claims the benefit of U.S. Provisional Application Serial No. 60/338,888 filed Nov. 30, 2001, the entire contents of which is incorporated herein by reference.

STATEMENT OF GOVERNMENT SUPPORT

[0002] This invention was made in part with government support under Grant No. CA65145 awarded by the National Institutes of Health. The government may have certain rights in this invention.

BACKGROUND OF THE INVENTION

[0003] 1. Field of the Invention

[0004] The present invention relates generally to methylation of genomic DNA and more specifically to the identification of sequences normally methylated in the genome and their relationship to disease states.

[0005] 2. Background Information

[0006] DNA methylation is central to many mammalian processes including embryonic development, X-inactivation, genomic imprinting, regulation of gene expression, and host defense against parasitic sequences, as well as abnormal processes such as carcinogenesis, fragile site expression, and cytosine to thymine transition mutations. DNA methylation in mammals is achieved by the transfer of a methyl group from S-adenosyl-methionine to the C5 position of cytosine. This reaction is catalyzed by DNA methyltransferases and is specific to cytosines in CpG dinucleotides. Seventy percent of all cytosines in CpG dinucleotides in the human genome are methylated and prone to deamination, resulting in a cytosine to thymine transition. This process leads to an overall reduction in the frequency of guanine and cytosine to about 40% of all nucleotides and a further reduction in the frequency of CpG dinucleotides to about a quarter of their expected frequency (Bird 1986).

[0007] The exception to CpG under representation in the genome is CpG islands, which were first identified as Hpa II tiny fragments (Bird et al. 1985), and were later formally defined as sequences >200 bp in length, with a GC content >0.5, and a CpGobs/CpGexp (observed to expected ratio based on GC content) >0.6 (Gardiner-Garden and Frommer 1987). CpG islands have been estimated to constitute 1%-2% of the mammalian genome (Antequera and Bird 1993), and are found in the promoters of all housekeeping genes, as well as in a less conserved position in 40% of genes showing tissue-specific expression (Larsen et al. 1992). The persistence of CpG dinucleotides in CpG islands is largely attributed to a general lack of methylation of CpG islands, regardless of expression status (reviewed in Cross and Bird 1995).

[0008] Although CpG islands are believed to be unmethylated, two exceptions to this rule in normal cells are the inactive X chromosome (Yen et al. 1984) and imprinted genes (Ferguson-Smith et al. 1993 ; Razin and Cedar 1994 ; Barlow 1995), both of which are associated with methylated CpG islands. Genomic imprinting is the parental origin-specific differential expression of the two alleles of a gene, and most imprinted genes show differential germline methylation of associated CpG islands (reviewed in Ohlsson et al. 2001). A third exception to the rule of methylation exclusion of CpG islands is aberrant methylation of CpG islands in tumors and in immortalized cultured cells, and such CpG island methylation is thought to contribute to carcinogenesis (Herman et al. 1994; Merlo et al. 1995).

[0009] Because of the interest in DNA methylation, genomic imprinting, and cancer, several general approaches have been used to identify CpG islands that are differentially methylated in specific cell types, such as screening tumor-normal pairs for cancer-related methylation changes (Huang et al. 1999; Shiraishi et al. 1999; Toyota et al. 1999), or pronuclear transplantation to examine differential parental origin for imprinted genes (Hayashizaki et. 1994 ; Plass et al. 1996). However, there are no reports of successfully using a systemic effort to identify unique, methylated CpG islands.

[0010] There are a variety of genome scanning methods that have been used to identify altered methylation sites in cancer cells. For example, one method involves restriction landmark genomic scanning (Kawai et al., Mol. Cell. Biol. 14:7421-7427, 1994), and another example involves methylation-sensitive arbitrarily primed PCR (Gonzalgo et al., Cancer Res. 57:594-599, 1997). Changes in methylation patterns at specific CpG sites have been monitored by digestion of genomic DNA with methylation-sensitive restriction enzymes followed by Southern analysis of the regions of interest. The digestion-Southern method is a straightforward method but it has inherent disadvantages in that it requires a large amount of DNA (at least or greater than 5 ug) and has a limited scope for analysis of CpG sites (as determined by the presence of recognition sites for methylation-sensitive restriction enzymes). Another method for analyzing changes in methylation patterns involves a PCR-based process that involves digestion of genomic DNA with methylation-sensitive restriction enzymes prior to PCR amplification (Singer-Sam et al., Nucl. Acids Res. 18:687, 1990). However, this method has not been shown effective because of a high degree of false positive signals (methylation present) due to inefficient enzyme digestion of over-amplification in a subsequent PCR reaction.

[0011] Genomic sequencing has been simplified for analysis of DNA methylation patterns and 5-methylcytosine distribution by using bisulfite treatment (Frommer et al., Proc. Natl. Acad. Sci. USA 89:1827-1831, 1992). Bisulfite treatment of DNA distinguishes methylated from unmethylated cytosines, but original bisulfite genomic sequencing requires large-scale sequencing of multiple plasmid clones to determine overall methylation patterns, which prevents this technique from being commercially useful for determining methylation patterns in any type of a routine diagnostic assay.

[0012] In addition, other techniques have been reported which utilize bisulfite treatment of DNA as a starting point for methylation analysis. These include methylation-specific PCR (MSP) (Herman et al. Proc. Natl. Acad. Sci. USA 93:9821-9826, 1992); and restriction enzyme digestion of PCR products amplified from bisulfite-converted DNA (Sadri and Hornsby, Nucl. Acids Res. 24:5058-5059, 1996; and Xiong and Laird, Nucl. Acids. Res. 25:2532-2534, 1997).

[0013] PCR techniques have been developed for detection of gene mutations (Kuppuswamy et al., Proc. Natl. Acad. Sci. USA 88:1143-1147, 1991) and quantitation of allelic-specific expression (Szabo and Mann, Genes Dev. 9:3097-3108, 1995; an d Singer-Sam et al., PCR Methods Appl. 1:160-163, 1992). Such techniques use internal primers, which anneal to a PCR-generated template and terminate immediately 5' of the single nucleotide to be assayed. However an allelic-specific expression technique has not been tried within the context of assaying for DNA methylation patterns.

[0014] Therefore, there remains a need for a method for using a systemic or genome-wide approach to identify unique, methylated CpG islands, GC rich regions and CpG dinucleotides, including normally methylated CpG sequences.

SUMMARY OF THE INVENTION

[0015] The present invention is based on the seminal discovery that normally methylated CpG islands or GC rich regions in the genome may lose methylation and this loss of methylation may be used to identify various diseases or disease states in a subject, imprinted genes and other characteristics of the genome.

[0016] In another aspect the present invention provides a method for identifying a CpG island or GC rich-regulated gene. It should be understood that while many of the illustrative examples in this invention show CpG islands, the invention includes not only CpG islands, but also GC rich regions and even CpG dinucleotide sequences. Thus, although the term island may be referred to, the term includes other GC rich sequences as well. The method includes identifying a candidate gene located on a chromosome near a CpG island and determining whether the expression of the candidate gene is regulated by methylation of the CpG island or GC rich region. In one illustrative example, the CpG island or GC rich regions used in the method include at least one of SEQ ID NO: 3-31. In certain embodiments, the method includes identifying the methylation state of a CpG island or GC rich region other than SEQ ID NO: 8 (gDMR 3-4), which has been identified as a gDMR (Arima et al. 2000).

[0017] In another aspect the present invention provides a method for identifying a population of CpG islands or GC rich regions in a genome. This aspect of the invention utilizes a method for isolating a library of normally methylated CpG island or GC rich regions disclosed herein. A method according to this aspect of the invention provides a genome-wide scan to identify a population of CpG islands or GC rich regions based on the combination of restriction enzymes used for the method. Therefore, a method according to this aspect of the invention identifies multi-copy CpG islands or GC rich regions within repeats as well as single copy CpG islands or GC rich regions. The method includes performing a double digestion by cleaving genomic DNA with both a restriction enzyme that cleaves at a recognition site with an AT content of greater than 50%, preferably greater than 75% AT, most preferably 100% AT, and a restriction endonuclease that cleaves at an unmethylated restriction site comprising greater than 50% CG, preferably greater than 75% GC, most preferably 100% GC, to generate a series of restriction fragments. The series of restriction fragments in length are typically size fractionated as discussed below, and fragments of a specified length (e.g. greater than 500 base pairs) are cloned in a restriction negative bacteria to generate a first library. This first cloning step enriches for CpG islands or GC rich regions and eliminates unmethylated CpG islands or GC rich regions because of the methylcytosine sensitivity of the restriction enzyme that recognizes only unmethylated restriction sites.

[0018] In another aspect, the present invention provides an isolated polynucleotide that includes a nucleotide sequence unmethylated in nucleic acid of paternal origin and methylated in nucleic acid of maternal origin. The polynucleotiode is about 1638 nucleotides encoding about 546 amino acids and has about 79% amino acid sequence identity to Elongin A2. In embodiment, the polynucleotide is set forth in SEQ ID NO: 1. This polynucleotide appears to be polymorphic at position 910, which can be G or A. In another embodiment, the polynucleotide encodes a polypeptide as set forth in SEQ ID NO: 2.

BRIEF DESCRIPTION OF THE DRAWINGS

[0019] **FIG. 1** illustrates an overall strategy for cloning methylated CpG islands. In step 1, genomic DNA is digested with Mse I which cuts between CpG islands, and Hpa II, which cuts unmethylated CpG islands. Mse I fragments containing methylated CpG islands then are transformed into a bacterial strain that does not cut methylated DNA. However, brief bacterial passage leads to loss of methylation of these previously methylated sequences. In step **2**, the library DNA is pooled and digested with Eag I, which cuts relatively large fragments within CpG islands, and these fragments are then subcloned.

[0020] **FIGS. 2A and 2B** illustrate methylation of CpG islands in normal human DNA. Genomic DNA from peripheral blood lymphocytes (A) or tissues (B) was digested with Mse I (M), Mse I+Hpa II (MH), or Mse I+Msp I (MM). Fragment sizes are indicated to the right. CpG islands used for Southern blot hybridization are indicated in panel A, and CpG island clone 1-19 was used in panel B. Note that there is an Mse I polymorphism in the fetal tissue that is not in the adult tissue, accounting for the presence of two bands in the fetal tissue Mse I digest. Blots were made in duplicate and one set was hybridized to RB to ensure the presence of DNA in the Msp I lane. BR, brain; CO, colon; KI, kidney; LI, liver; fCNS, fetal CNS; fKI, fetal kidney; ELU, fetal lung; fSK, fetal skin.

[0021] **FIGS. 3A and 3B** show a series of gels showing differential methylation of novel gDMRs in uniparental tissues of germline origin. Fragment sizes (kb) are indicated to the right. (A) Sperm (SP), ovarian teratoma (OT), or complete hydatidiform mole (CHM) was digested, and Southern blot hybridization was performed with the gDMRs indicated, as described in the legend to **FIG. 2**. Multiple OT and CHM were examined with similar results, although only one is shown. (B) Similar experiments were performed with an unmethylated CpG island in the retinoblastoma gene (RB), with a CpG island upstream of H19 that shows preferential methylation of the paternal allele, and with a CpG island within the SNRPN gene that shows preferential methylation of the maternal allele.

[0022] **FIG. 4** shows a series of gels showing similar methylation of novel SMRs in uniparental tissues of germ-

line origin. Experiments were performed as described in the legend to **FIG. 2**, using the SMRs indicated. Fragment sizes are indicated to the right.

[**0023**] **FIG. 5** illustrates the chromosomal location and relationship of representative methylated CpG islands to nearby genes. Genes are indicated with boxes, and the arrows show transcriptional orientation. The methylated CpG islands are shown in shading. In the case of 2-78, the homologous sequence within Elongin A2 is indicated.

[**0024**] **FIG. 6** shows the nucleotide and amino acid sequence of Elongin A3 (SEQ ID NO: 1 and 2, respectively). The transcription factor SII similarity motif is shown by the boldfaced bases in the top 6 lines of the figure. The nuclear localization signal is shown by the boldfaced bases in the bottom 2 lines of the figure. The site of the (G/A) polymorphism used for imprinting analysis is boldfaced at nucleotide 910, and the PCR primers specific for Elongin A3 are shown in boldfaced type beginning on the lines that have number 811 and 1261 to the left.

[**0025**] FIGS. **7**A-D illustrates tissue-specific imprinting of Elongin A3. The (G/A) polymorphism was used to assess allele-specific expression in four heterozygous fetuses denoted A, B, C, and D. Chromatograms of genomic DNA (gDNA) sequence are included to show heterozygosity, as well as the homozygous maternal decidual DNA indicating parental origin. (A) Monoallelic expression of the maternal allele in lung, central nervous system (CNS), and limbs, and biallelic expression in kidney. (B) Monoallelic expression of the maternal allele in placenta and CNS, and biallelic expression in intestine. (C) Monoallelic expression of the maternal allele in CNS, biallelic expression in kidney and liver. (D) Monoallelic expression of the maternal allele in placenta, and biallelic expression in kidney and liver. Sequencing was done bidirectionally in all cases, and monoallelic expression of the maternal allele did not depend on whether that allele was A or G.

[**0026**] **FIG. 8** shows sequence conservation of methylated CpG islands between human and mouse. Human methylated CpG islands and ~1 kb of flanking DNA were compared to mouse sequence, synteny was confirmed, the corresponding mouse CpG islands were identified, and regions of conservation (percentage shown) were determined. In the case of gDMR 1-21, the corresponding mouse sequence, while GC-rich, showed an observed to expected CpG ratio of 0.45-0.50 and therefore was not classified as a CpG island.

DETAILED DESCRIPTION OF THE INVENTION

[**0027**] To identify chromosomal regions that might harbor imprinted genes, the present invention provides a method for generating a library of normally methylated GC rich regions (e.g., a CpG island). Most of the nucleic acid sequences containing methylated CpG islands or GC rich regions isolated using the methods of the invention are high copy number dispersed repeats. However, unique clones in the library can be identified and characterized. Some of the unique clones identified herein were differentially methylated in uniparental tissue of germline origin. These clones are referred to herein as germline differentially methylated regions (gDMRs).

[**0028**] Surprisingly, many of the methylated CpG islands or GC rich regions identified in the Examples herein, are methylated in germline tissues of both parental origins, representing a previously uncharacterized class of normally methylated CpG islands or GC rich regions in the genome, and which we term similarly methylated regions (SMRs). These SMRs, in contrast to the gDMRs, are shown herein to be significantly associated with telomeric band locations, suggesting a potential role for SMRs in chromosome organization. Finally, many of the methylated CpG islands or GC rich regions are on average 85% conserved between mouse and human. While many CpG or GC rich regions are CpG islands, the methods of the invention are not limited to CpG islands.

[**0029**] In one embodiment, the invention provides a method for determining a disease state in a subject by determining the DNA methylation status at a cytosine residue of a CpG sequence in a genomic DNA sample from the subject, wherein hypomethylation of a CpG sequence normally methylated in a subject not having the disease state, is indicative of a disease state in the subject. The CpG sequence is typically found within a GC rich region or a CpG island. The invention methods are preferably used when the subject is a human. Although the disease state is often cancer, the invention is not so limited. The disease state includes other diseases such as multiple sclerosis, Alzheimer's disease, Parkinson's disease, depression and other imbalances of mental stability, atherosclerosis, cystic fibrosis, diabetes, obesity, Crohn's disease, and altered circadian rhythmicity, arthritis, inflammatory reactions or disorders, psoriasis and other skin diseases, autoimmune diseases, allergies, hypertension, anxiety disorders, schizophrenia and other psychoses, osteoporosis, muscular dystrophy, amyotrophic lateral sclerosis or circadian rhythm-related conditions.

[**0030**] In another embodiment, the invention provides a method for determining the DNA methylation status at a cytosine residue of a CpG sequence in a genomic DNA sample by performing methylation state analysis of one or more CpG islands or GC rich regions of a genomic DNA sample, thereby determining the DNA methylation status in the genomic DNA sample. One method for performing methylation state analysis is exemplifed in the Examples herein. In one aspect, the one or more CpG islands or GC rich regions include differentially methylated regions (DMRs). In another aspect, the one or more CpG islands and GC rich regions include similarly methylated regions (SMRs).

[**0031**] In one aspect the present invention provides a method for identifying a CpG island or GC rich region methylation state that includes performing methylation state analysis of one or more CpG islands or GC rich regions of a genomic DNA sample, wherein the one or more CpG islands or GC rich regions are at least one of SEQ ID NOs: 1-31, and in certain embodiments, SEQ ID NOs: 3-7 and 9-31.

[**0032**] In one aspect the present invention provides a method for identifying a CpG island or GC rich region methylation state that includes performing methylation state analysis of one or more CpG islands or GC rich regions of a genomic DNA sample. In one aspect, the one or more CpG islands or GC rich regions are at least one of SEQ ID NOs: 1-31, and in certain embodiments, SEQ ID NOs: 3-7 and 9-31 with the proviso that it is not SEQ ID NO: 8. CpG

islands or GC rich regions are sequences greater than 200 bp in length, with a GC content >0.5, and a CpGobs/CpGexp (observed to expected ratio based on GC content) >0.6 (Gardiner-Garden and Frommer 1987).

[0033] These methods are useful in providing information regarding gene regulation since it is known that methylation of CpG islands or GC rich regions affects gene expression (Ferguson-Smith et al. 1993; Razin and Cedar 1994; Barlow 1995, Ohlsson et al. 2001, Herman et al. 1994; and Merlo et al. 1995). For example, expression of a tumor suppressor gene can be abolished by de novo DNA methylation of a normally unmethylated CpG island or GC rich region (Issa, et al., Nature Genet., 7:536, 1994; Herman, et al., supra; Merlo, et al., Nature Med., 1:686, 1995; Herman, et al., Cancer Res., 56:722, 1996; Graff, et al., Cancer Res., 55:5195, 1995; Herman, et al., Cancer Res., 55:4525, 1995). Consistent with the role of the CpG islands or GC rich regions identified herein in gene regulation, most of the methylated CpG islands or GC rich regions disclosed herein are localized within or near the coding sequence of known genes or of anonymous ESTs within the GenBank or Celera databases. The GC rich regions may be in exons, introns or regulatory regions, for example.

[0034] In all the methods described herein, the identification of sequences normally methylated and which have lost methylation is used for identifying a disease or disease state. Such disease or disease state includes cancer, multiple sclerosis, Alzheimer's disease, Parkinson's disease, depression and other imbalances of mental stability, atherosclerosis, cystic fibrosis, diabetes, obesity, Crohn's disease, and altered circadian rhythmicity, arthritis, inflammatory reactions or disorders, psoriasis and other skin diseases, autoimmune diseases, allergies, hypertension, anxiety disorders, schizophrenia and other psychoses, osteoporosis, muscular dystrophy, amyotrophic lateral sclerosis and circadian rhythm-related conditions. Preferred subjects for the present methods are mammals such as humans.

[0035] The methylation state of CpG refers to whether a particular cytidine residue in a CpG containing dinucleotide contains any degree of methylation. A CpG dinucleotide or a CpG island or GC rich region is characterized as either methylated or non-methylated based on whether any cytidines of the CpG island or GC rich region are methylated. The methylation state of a CpG island or GC rich region may be completely unmethylated, completely methylated, or partially methylated, and the degree of methylation can be quantified as a percent of residues methylated, as well as individually methylated CpG sites identified. In addition, a particular site can be variably methylated in a population of cells, and that degree of methylation can be quantified. Prior to the present invention, it had been thought that CpG dinucleotides or islands or GC rich regions were typically unmethylated, meaning that the degree of methylation would be nearly zero or quite low (such as less than 10%). Thus, a degree of "normal" methylation greater than the nearly zero amount would be refered to as a "normally" methylated CpG dinucleotide.

[0036] Methylation state analysis of CpG islands or GC rich regions can be performed by any method known in the art. Most of the methods developed to date for detection of methylated cytosine depend upon cleavage of the phosphodiester bond alongside cytosine residues, using either methylation-sensitive restriction enzymes or reactive chemicals such as hydrazine which differentiate between cytosine and its 5-methyl derivative. Examples of methylation sensitive restriction endonucleases which can be used to detect 5'CpG methylation include SmaI, SacII, EagI, MspI, HpaII, BstUI and BssHII, for example.

[0037] Genomic sequencing protocols which identify a 5-MeC residue in genomic DNA as a site that is not cleaved by any of the Maxim Gilbert sequencing reactions can also been used. Other techniques utilize bisulfite treatment of DNA as a starting point for methylation analysis. These include methylation-specific PCR (MSP) (Herman et al. Proc. Natl. Acad. Sci. USA 93:9821-9826, 1992); and restriction enzyme digestion of PCR products amplified from bisulfite-converted DNA (Sadri and Hornsby, Nucl. Acids Res. 24:5058-5059, 1996; and Xiong and Laird, Nucl. Acids. Res. 25:2532-2534, 1997). See also 6,262,171 6,200, 756 6,017,704 5,786,146, all incorporated herein by reference.

[0038] In certain embodiments of this aspect of the present invention, CpG island or GC rich region methylation state is determined for similarly methylated regions (SMRs). The Examples included herein utilize methods of the present invention to identify numerous human SMRs. SMRs are CpG islands or GC rich regions that are methylated equally in male and female tissue of germline origin. The SMRs can include at least one of SEQ ID NO: 3, SEQ ID NO: 4, SEQ ID NO: 5, SEQ ID NO: 6, SEQ ID NO: 7, SEQ ID NO: 9, SEQ ID NO: 10, SEQ ID NO: 11, SEQ ID NO: 12, SEQ ID NO: 16, SEQ ID NO: 21, SEQ ID NO: 22, SEQ ID NO: 23, SEQ ID NO: 26, SEQ ID NO: 29, and SEQ ID NO: 30, as identified in Table 1. Thus, in one aspect, the invention provides a method for determining the methylation status of a population of similarly methylated regions (SMRs) in a subject by performing methylation status analysis of a population of SMRs of genomic DNA from a human sample. In one aspect, the methylation status of SMRs is correlated with a disease state. In one aspect, the population of SMRs comprises at least two SMRs and in one aspect, at least three SMRs.

[0039] Of the sixteen SMRs identified herein, sixteen of seventeen were localized near the ends of chromosomes, either on the last (n=15) or the penultimate (n=1) subband of the chromosome on which it resides. The method of this aspect of the invention can identify the methylation state of an SMR located near the end of a chromosome. Table 2 and **FIG. 5** show the location of specific SMRs of the invention. The Examples included herein also identify CpG islands or GC rich regions that are differentially methylated in germline tissue of male and female origin. These CpGs are referred to herein as germline differentially methylated regions (gDMRs). The method of this aspect of the invention can identify the methylation state of a gDMR. For example, the methylation state can be determined for are least one of SEQ ID NO: 13, SEQ ID NO: 14, SEQ ID NO: 15, SEQ ID NO: 17, SEQ ID NO: 19, SEQ ID NO: 20, SEQ ID NO: 24, SEQ ID NO: 25, SEQ ID NO: 27, SEQ ID NO: 28, and SEQ ID NO: 31.

[0040] The methylated CpG islands or GC rich regions identified herein were distributed throughout the genome. There was a striking localization of SMRs near the ends of chromosomes. Sixteen of 17 SMRs were localized near the

ends of chromosomes, either on the last (n=15) or the penultimate (n=1) subband of the chromosome on which it resided (Table 2). In contrast, of 12 gDMRs that could be mapped (of the 13 gDMRs studied), only four were localized near the ends of chromosomes (Table 2). This difference was highly statistically significant (P=0.0008, Fisher's exact test). The association of SMRs near the ends of chromosomes is consistent with an observation of densely methylated GC-rich sequences near telomeres, although that study did not describe methylated CpG islands or GC rich regions (Brock et al. 1999). In addition, there was a segregation of gDMRs and SMRs within compartments of differing genomic composition, i.e., isochores, which are regions of several hundred kilobases of relatively homogeneous GC composition (Bernardi 1995). Approximately 75% of the SMRs fell within high isochore regions (G+C 50%), as might be expected from the high GC content of methylated CpG islands or GC rich regions. Surprisingly, however, all of the gDMRs fell within low isochore regions (G+C<50%), i.e., of relatively low GC content, despite the high GC content of the gDMRs themselves (L. Z. Strichman-Almashanu and A. P. Feinberg). This difference was statistically significant (P<0.01, Fisher's exact test). Thus, the gDMRs and SMRs may lie within distinct chromosomal and/or isochore compartments. These results provide the basis for a method to identify epigenetic chromosomal domains. Localization of CpG islands or GC rich regions to the telo/subtelo regions, for example, can be used for identifying imprinted gene domains, disease domains (e.g. p16), chromatin regulated genes controlled at a distance, such as telomerase (TERT) or c-myc by CTCF; and developmentally programmed regions essential for organ formation, such as the brain in Lunyak et al. (Science. Oct. 24, 2002), for example.

[0041] The method of this aspect of the invention for identifying the CpG island or GC rich region methylation state can involve identifying the methylation state of one, or more than one CpG island or GC rich region. For example, the methylation state of 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 20, 25 CpG islands or GC rich regions, or in certain embodiments all 29 CpG islands or GC rich regions disclosed herein (SEQ ID NOs:3-31 is identified). In fact, according to the present invention the methylation state of any of the CpG islands or GC rich regions disclosed in Table 1 can be determined.

[0042] In a embodiment of this aspect of the present invention, the methylation state of SEQ ID NO: 27 is identified. gDMR1-13 (SEQ ID NO: 27) is located on 18q23 within a predicted gene of unknown function, and near the SALL3 gene, a candidate gene for 18q deletion syndrome, which involves preferential loss of the paternal allele (Kohlhase et al. 1999).

[0043] In another embodiment of this aspect of the present invention, the methylation state of SEQ ID NO: 30 is identified. SMR1-2 (SEQ ID NO: 30) is located on 19q13.4 within 110 kb of a glioma tumor suppressor candidate gene.

[0044] In another embodiment of this aspect of the present invention, the method includes identifying the methylation state of SEQ ID NO: 4 (SMR 3-20). This CpG island or GC rich region is located within the HDAC4 gene (See **FIG. 5**) and there are several other predicted genes and antisense transcripts near this CpG island or GC rich region.

[0045] In another embodiment of this aspect of the present invention, the method includes identifying the methylation state of SEQ ID NO: 26, located within 16 kb from CpG island or GC rich region 2-3.

[0046] In another embodiment of this aspect of the present invention, the method includes identifying the methylation state of SEQ ID NO: 21. SEQ ID NO: 21 (SMR 3-110) is located near a predicted apoptosis inhibitor, a septin-like cell division gene, a ras homolog, and a predicted translation initiation factor.

[0047] In another embodiment of this aspect of the present invention, the method includes identifying the methylation state of SEQ ID NO: 23. SEQ ID NO: 23 (SMR 1-12) is located near a predicted apoptosis inhibitor, a septin-like cell division gene, a ras homolog, and a predicted translation initiation factor.

[0048] In another embodiment of this aspect of the present invention, the method includes identifying the methylation state of SEQ ID NO: 21 (SMR 3-110) and SEQ ID NO: 23 (SMR 1-12). Together these CpGs flank a predicted apoptosis inhibitor, a septin-like cell division gene, a ras homolog, and a predicted translation initiation factor.

[0049] In another aspect, the present invention provides a method for determining the methylation state of a series of similarly methylated regions (SMRs) in a subject, the method comprising performing methylation state analysis on a series of SMRs of genomic DNA from a human sample. The present disclosure reveals that the presence of normally-methylated single-copy CpG islands or GC rich regions are more abundant than previously believed. The ability to analyze the methylation state of a series of these normally methylated CpGs provides valuable information regarding the overall methylation state of a genome. This information may provide information regarding overall chromatin state of a genome since SMRs appear to be located near the ends of chromosomes, as illustrated in the Examples herein. Furthermore, such information may provide prognostic, diagnostic, or disease monitoring tools related to cancer, based on previous observations that implicate methylation of genomic methylation in cancer.

[0050] The series of SMRs whose methylation state is determined can include at least two, three, four, five, ten, 15, or all of SEQ ID NO: 3, SEQ ID NO: 4, SEQ ID NO: 5, SEQ ID NO: 6, SEQ ID NO: 7, SEQ ID NO: 9, SEQ ID NO: 10, SEQ ID NO: 1 1, SEQ ID NO:12, SEQ ID NO: 16, SEQ ID NO: 21, SEQ ID NO: 22, SEQ ID NO: 23, SEQ ID NO: 26, SEQ ID NO: 29, and SEQ ID NO: 30.

[0051] As discussed above, methods of the present invention for identifying a CpG island or GC rich region methylation state and methods below for identifying a population of CpG islands or GC rich regions, provide valuable information regarding gene regulation since it is known that methylation of CpG islands or GC rich regions can affect expression of genes located near the CpG island or GC rich region. Accordingly, in one aspect, the present invention provides a method for identifying the presence of an imprinted gene that includes comparing the methylome of genomic DNA of maternal origin with the methylome of genomic DNA of paternal origin, wherein a difference in methylation patterns between the two methylomes is indicative of the presence of an imprinted gene. A methylome is

the methylation pattern of an entire genome (Feinberg 2001). DNA methylation serves as an additional layer of genetic information in the genome (Feinberg 2001). Typically, methylation in a genome occurs in CpG islands or GC rich regions. A methylome of a subject can be determined using methods disclosed herein.

[0052] The present invention includes an imprinted gene identified by the above method. Genomic imprinting is the parental origin-specific differential expression of the two alleles of a gene. Most imprinted genes show differential germline methylation of associated CpG islands or GC rich regions (reviewed in Ohlsson et al. 2001).

[0053] In another embodiment of this aspect of the invention, a method is provided for identifying the presence of an imprinted gene, that includes identifying a population of CpG islands or GC rich regions and identifying a candidate gene found within 200 kilobases of a first CpG island or GC rich region of the population of CpG islands or GC rich regions. A determination is made of whether the candidate gene is regulated by methylation of the first CpG rich region of the population of CpG islands or GC rich regions and preferentially methylated in genomic DNA from paternal or maternal origin. Regulation of the candidate gene by methylation of the first CpG island or GC rich region and paternal or maternal preferential methylation is indicative of an imprinted gene. The present invention includes imprinted genes identified by the above method.

[0054] In certain embodiments, the first CpG island or GC rich region is gDMR 3-4 (SEQ ID NO: 27). Interestingly, gDMR 3-4 is located on 18q23, which has been implicated in bipolar affective disorder, specifically harboring a predisposing gene transmitted preferentially through the father (Stine et al. 1995 ; McMahon et al. 1997). Therefore, the localization of this gDMR herein can serve as a guidepost for identifying candidate imprinted genes for this important disease.

[0055] In another aspect the present invention provides a method for identifying a CpG island or GC rich region-regulated gene. The method includes identifying a candidate gene located on a chromosome near a CpG island or GC rich region and determining whether the expression of the candidate gene is regulated by methylation of the CpG island or GC rich region. Preferably, the CpG islands or GC rich regions used in the method include at least one of SEQ ID NO: 3-31. In certain embodiments, the method includes identifying the methylation state of a CpG island or GC rich region other than SEQ ID NO: 8 (gDMR 3-4), which has been identified as a gDMR (Arima et al. 2000).

[0056] A CpG island or GC rich region-regulated gene is a gene whose expression is regulated by methylation of a CpG island or GC rich region. A CpG island or GC rich region is located near a candidate gene when it is located within about 2000, 1000, 500, 200 or 100 kilobases of the gene. In other embodiments of the invention, the gene is located within about 50, 25, 10, 5, or 1 kilobase of the gene. In other embodiments, the CpG island or GC rich region is located within a candidate gene. For example, Prader-Willi syndrome CpG island or GC rich region in exon 1 of SNRPN controls expression of genes up to 2 megabases away, e.g. Buiting et al. Inherited microdeletions in the Angelman and Prader-Willi syndromes define an imprinting centre on human chromosome 15. Nat Genet. April 1995; 9(4):395-400.

[0057] A determination of genes on the same chromosome as a CpG island or GC rich region, and the approximate distance between a CpG island or GC rich region and a candidate gene can be determined by mapping the CpG island or GC rich region and candidate gene sequences using human genome sequence information on databases such as GenBank (available at http://www.ncbi.nlm.nih.gov/) or the Celera human gene sequence database.

[0058] Methods are known in the art for determining whether expression of a candidate gene is regulated by methylation of a CpG island (see e.g., Ferguson-Smith et al. 1993; Razin and Cedar 1994; Barlow 1995, Ohlsson et al. 2001, Herman et al. 1994; and Merlo et al. 1995). For example, the effect on gene expression of de novo DNA methylation of a normally unmethylated CpG island, can be analyzed (Issa, et al., Nature Genet., 7:536, 1994; Herman, et al., supra; Merlo, et al., Nature Med., 1:686, 1995; Herman, et al., Cancer Res., 56:722, 1996; Graff, et al., Cancer Res., 55:5195, 1995; Herman, et al., Cancer Res., 55:4525, 1995).

[0059] In one embodiment, the method includes determining whether the candidate gene is regulated by methylation of SEQ ID NO: 13, SEQ ID NO: 14, SEQ ID NO: 15, SEQ ID NO: 17, SEQ ID NO: 19, SEQ ID NO: 20, SEQ ID NO: 24, SEQ ID NO: 25, SEQ ID NO: 27, SEQ ID NO: 28, and SEQ ID NO: 31. This embodiment, includes the CpGs identified herein as being gDMRs.

[0060] In a embodiment of this aspect of the present invention, the method includes determining whether the candidate gene is regulated by methylation of SEQ ID NO: 27. gDMR1-13 (SEQ ID NO: 27) is located on 18q23 within a predicted gene of unknown function, and near the SALL3 gene, a candidate gene for 18q deletion syndrome, which involves preferential loss of the paternal allele (Kohlhase et al. 1999).

[0061] In another embodiment of this aspect of the present invention, the method includes determining whether the candidate gene is regulated by methylation of SEQ ID NO:30. SMR1-2 (SEQ ID NO: 30) is located on 19q13.4 within 110 kb of a glioma tumor suppressor candidate gene.

[0062] In another embodiment of this aspect of the present invention, the method includes determining whether the candidate gene is regulated by methylation of SEQ ID NO: 4 (SMR 3-20). This CpG island or GC rich region is located within the HDAC4 gene (See **FIG. 5**) and there are several other predicted genes and antisense transcripts near this CpG island or GC rich region.

[0063] In another embodiment of this aspect of the present invention, the method includes determining whether the candidate gene is regulated by methylation of SEQ ID NO: 26, located within 16 kb from CpG island or GC rich region 2-3.

[0064] In another embodiment of this aspect of the present invention, the method includes determining whether the candidate gene is regulated by methylation of SEQ ID NO:21. SEQ ID NO: 21 (SMR 3-110) is located near a predicted apoptosis inhibitor, a septin-like cell division gene, a ras homolog, and a predicted translation initiation factor.

[0065] In another embodiment of this aspect of the present invention, the method includes determining whether the

candidate gene is regulated by methylation of SEQ ID NO: 23. SEQ ID NO: 23 (SMR 1-12) is located near a predicted apoptosis inhibitor, a septin-like cell division gene, a ras homolog, and a predicted translation initiation factor.

[0066] In another embodiment of this aspect of the present invention, the method includes determining whether the candidate gene is regulated by methylation of SEQ ID NO: 21 (SMR 3-110) and SEQ ID NO: 23 (SMR 1-12). Together these CpGs flank a predicted apoptosis inhibitor, a septin-like cell division gene, a ras homolog, and a predicted translation initiation factor.

[0067] In another embodiment of this aspect of the present invention, the method includes determining whether the candidate gene is regulated by methylation of a similarly methylated region (SMR). For example, the method can determine whether the candidate gene is regulated by methylation of SEQ ID NO: 3, SEQ ID NO: 4, SEQ ID NO: 5, SEQ ID NO: 6, SEQ ID NO: 7, SEQ ID NO: 9, SEQ ID NO: 10, SEQ ID NO: 11, SEQ ID NO: 12, SEQ ID NO: 16, SEQ ID NO: 21, SEQ ID NO: 22, SEQ ID NO: 23, SEQ ID NO: 26, SEQ ID NO: 29, or SEQ ID NO: 30.

[0068] In another aspect the present invention provides a method for identifying a population of CpG islands or GC rich regions in a genome. In one illustrative aspect, the invention utilizes a method for isolating a library of normally methylated CpG islands or GC rich regions disclosed herein. A method according to this aspect of the invention provides a genome-wide scan to identify a population of CpG islands or GC rich regions based on the combination of restriction enzymes used for the method. Therefore, a method according to this aspect of the invention identifies multi-copy CpG islands or GC rich regions within repeats as well as single copy CpG islands or GC rich regions. The method includes performing a double digestion by cleaving genomic DNA with both a restriction enzyme that cleaves at a recognition site with an AT content of greater than 50%, preferably greater than 75% AT, most preferably 100% AT, and a restriction endonuclease that cleaves at an unmethylated restriction site comprising greater than 50% CG, preferably greater than 75% GC, most preferably 100% GC, to generate a series of restriction fragments. The series of restriction fragments in length are typically size fractionated as discussed below, and fragments of a specified length (e.g. greater than 500 base pairs) are cloned in a restriction negative bacteria to generate a first library. This first cloning step enriches for CpG islands or GC rich regions and eliminates unmethylated CpG islands or GC rich regions because of the methylcytosine sensitivity of the restriction enzyme that recognizes only unmethylated restriction sites.

[0069] The next step (i.e. the second cloning step) provides further enrichment of CpG islands or GC rich regions by digesting DNA from the first library with an infrequently cutting restriction endonuclease specific for sequences common to GC islands or GC rich regions (e.g., a CpG rich region infrequent restriction endonuclease). An infrequently cutting restriction endonuclease is an endonuclease that recognizes a GC-rich recognition site (e.g., greater than 50% GC content) of at least 6 base pairs in length. (see Gardiner-Garden and Frommer, 1987). As used herein, the methods of the invention include not only CpG islands or GC rich regions (e.g., greater than 200 bp in length and a GC content of >0.5) or GC-islands or GC rich regions in which CpGobs/CpGexp >0.6, but also CpG islands or GC rich regions that do not meet these threshhold requirements but which are GC rich and contain multiple CpG dinucleotides. (see also Strichman-Almashanu et al., 2002, herein incorporated by reference in its entirety). Preferably the recognition site recognized by the infrequently cutting restriction endonuclease is has a GC content of at least 75%, most preferably 100% GC. This second cloning step results in isolation of relatively large fragments of CpG islands or GC rich regions that are normally methylated (i.e., survived the first cloning step), but are now unmethylated in the library and therefore amenable to digestion and subcloning.

[0070] Virtually any endonuclease that cleaves at a restriction site with a GC content of at least 50%, preferably 75%, and most preferably 100% can be used for the first cloning step in combination with virtually any endonuclease that cleaves at a restriction site with an AT content of at least 50%, preferably 75%, and most preferably 100%. For example, the GC-rich recognition site cleaving enzyme include but are not limited to HpaII, BtgI, SacII, NgoM IV, Bssh II, NaeI, Eag I, BsiE I, Kas I, PspOM I, NarI, SfoI, or Apa I. The AT-rich recognition site cleaving enzyme can be, for example, MseI, SspI, DraI, Tsp509I, ApoI, SspI, AseI, PsiI, DraI. In one embodiment, a double digestion is performed with Mse I, which recognizes the sequence TAAA and Hpa II, which recognizes the sequence CCGG at unmethylated sites.

[0071] The CG-rich recognition site cleaving enzyme and the AT-rich recognition site cleaving enzyme can be used in any order or simultaneously depending on the required reaction conditions for the restriction enzymes used, as will be understood.

[0072] A restriction negative bacteria is used in the first cloning step in order to avoid bacterial digestion of methylated genomic DNA. Virtually any restriction negative bacteria can be used in the methods of the present invention. For example the restriction negative bacterium can be XL2-Blue MRF'. Many strains of bacteria have been derived that are deficient in, for example, the bacterial enzymes mcrA, mcrCB and mrr. Another example is Strategene XL10 Gold. One bacterium for the first cloning step is the restriction-negative strain XL2-Blue MRF' to avoid bacterial digestion of methylated genomic DNA.

[0073] For the restriction digest before the second cloning step, virtually any endonuclease that recognizes a GC-rich recognition site at least 6 base pairs in length can be used. Examples of restriction endonucleases that can be used in this step include Eag 1. In certain embodiments, the restriction endonuclease Eag 1 (recognition sequences CGGCCG) is used. In these embodiments using Eag 1, the resulting library can be referred to as the Eag library.

[0074] Preferably, after the digestions before the first and second cloning steps, DNA fragments of specified lengths are isolated and cloned. For example, fragments of at least 100 bp, 250 bp, 500 bp, 2500 bp, and in certain embodiments at least 1000 bp are isolated and cloned. In other embodiments, DNA fragments of specified size ranges can be isolated and cloned. For example, fragments of 100-500 bp, 500-1000 bp, and greater than 1000 bp can be isolated and cloned separately. Methods are well known in the art for size fractionating nucleic acids, such as by using gel purification.

[0075] By repeating the aforementioned method for identifying a population of CpG islands or GC rich regions in a genome with different restriction enzymes, and by utilizing various known methods such as methylation specific PCR or bisulfite methods, described herein and known in the art, virtually the entire methylome of an organism can be determined. Alternatively, the method for identifying a population of low copy number CpG islands or GC rich regions can be repeated with different restriction enzymes to identify virtually all the low copy number CpG islands or GC rich regions of a methylome.

[0076] The population of CpG islands or GC rich regions in methods of this aspect of the invention can include a subset of least about 50, 100, 200, 250, 500, or 1000 palindromic CpG sites. Additionally, the population of CpG islands or GC rich regions can include at least about 2, 3, 4, 5, 10, 20, 25, 50, or 100 distinct CpG islands or GC rich regions.

[0077] The ability to characterize entire methylomes provides further uses for the methods of the invention. For example, methylomes of the same species, for example human methylomes, or portions of a methylome identified using a first set of restriction enzymes to perform the above method of the invention, can be compared to identify methylation differences that are involved in phenotypic differences among individuals of a species. Furthermore, methylomes, or portions thereof, between species can be compared to identify CpG islands or GC rich regions that are important gene expression regulators, by identifying CpG islands or GC rich regions that are conserved between species. The Examples included herein provide a comparison of portions of the methylome of mouse and man to identify conserved CpG islands or GC rich regions.

[0078] Furthermore, the method discussed above for identifying a population of CpG islands or GC rich regions in a genome, can be used to identify the methylation state of a series of CpG islands or GC rich regions in various tissues and to determine whether methylation of a CpG island or GC rich region is preferentially related to cells from one parent, or certain tissues, as illustrated in the Examples provided herein. As illustrated in the Examples section hereinbelow, 62 unique CpG island or GC rich region clones were isolated and characterized using methods of the present invention, all of which were methylated and GC-rich, with a GC content >50%. Of these, 43 clones also showed a CpGobs/CpGexp >0.6, of which 30 were studied in detail. These unique methylated CpG islands or GC rich regions mapped to 23 chromosomal regions, and 12 were differentially methylated regions in uniparental tissues of germline origin, i.e., hydatidiform moles (paternal origin) and complete ovarian teratomas (maternal origin), even though many apparently were methylated in somatic tissues. At least two gDMRs mapped near imprinted genes, HYMA1 and a novel homolog of Elongin A and Elongin A2, which we term Elongin A3 (NM_145653), discussed in further detail below. Surprisingly, 18 of the methylated CpG islands or GC rich regions were methylated in germline tissues of both parental origins, representing a previously uncharacterized class of normally methylated CpG islands or GC rich regions in the genome, referred to herein as similarly methylated regions (SMRs). These SMRs, in contrast to the gDMRs, were significantly associated with telomeric band locations (P=0.0008), suggesting a potential role for SMRs in chromosome organization. At least 10 of the methylated CpG islands or GC rich regions identified herein were on average 85% conserved between mouse and human. These sequences will provide a valuable resource in the search for novel imprinted genes, for defining the molecular substrates of the normal methylome, and for identifying novel targets for mammalian chromatin formation.

[0079] Evidence for loss of methylation in cancer, (Feinberg and Vogelstein, 1983) has been shown by hypomethylation in genes of some human cancers as compared to their normal counterparts (Nature. Jan. 6, 1983;301(5895):89-92). More recently, this has been shown in the activation of MAGE melanoma antigen (Serrano, A., Garcia, A., Abril, E., Garrido, F., and Ruiz-Cabello, F. 1996). Methylated CpG points identified within MAGE-1 promoter were shown to be involved in gene repression. (Int. J. Cancer 68: 464-470 and De Smet, C., De Backer, O., Faraoni, I., Lurquin, C., Brasseur, F., and Boon, T. 1996). The activation of human gene MAGE-1 in tumor cells was correlated with genome-wide demethylation (Proc. Natl. Acad. Sci. 93: 7149-7153).

[0080] In another aspect the present invention provides a method for identifying a population of low copy number CpG islands or GC rich regions. A method according to this aspect of the invention includes cleaving genomic DNA with both a restriction enzyme that cleaves at a recognition site comprising adenosine and thymidine residues and a restriction endonuclease that cleaves at an unmethylated restriction site comprising cytidine and guanosine residues, to generate a series of restriction fragments and excluding those that are methylated, and cloning restriction fragments of at least 200, 300, 400, 500 and the like kb from the series of restriction fragments in a restriction negative bacteria to generate a first library. This step is similar to the initial double digestion and first cloning step discussed above for the aforementioned aspect of the invention. (See **FIG. 1**).

[0081] After generating the first library, the cloned DNA of the first library is cleaved with a restriction enzyme that cleaves DNA at a restriction site within a CpG island or GC rich region; excluding CpG island or GC rich region fragments that contain repetitive elements while leaving low copy CpG island or GC rich region fragments intact, thereby producing a population of low copy number CpG islands and GC rich region fragments. Such fragments may be at least about 100, 200, 300, 400, 500 and the like kb in size. The method may further include cloning the restriction fragments containing low copy CpG islands and GC rich regions to form a library containing a plurality of low copy CpG island or GC rich region DNA. The "excluding" step is by optionally cleaving cloned DNA of the first library with a restriction enzyme that cleaves DNA at a restriction site within a CpG island or GC rich region repeat sequence or using a methylated CpG binding column, or other methods known to those of skill in the art. A final library containing a plurality of clones is also included in the invention. In one aspect, the GC rich regions are CpG islands.

[0082] In another aspect, the present invention provides an isolated polynucleotide that includes a nucleotide sequence unmethylated in nucleic acid of paternal origin and methylated in nucleic acid of maternal origin. The polynucleotiode is about 1638 nucleotides encoding about 546 amino acids and has about 79% amino acid sequence identity to Elongin A2. In embodiment, the polynucleotide is set forth in SEQ

ID NO: 1. This polynucleotide appears to be polymorphic at position 910, which can be G or A. In another embodiment, the polynucleotide encodes a polypeptide as set forth in SEQ ID NO: 2.

[0083] The polynucleotide of SEQ ID NO: 1 is a novel imprinted gene that was identified using methods of the present invention (as illustrated in the Examples below). The CpG island or GC rich region gDMR 2-78 was localized to 18q21 (**FIG. 5**) and was completely methylated in all somatic fetal and adult tissues tested (**FIG. 2**). However, this CpG rich region was unmethylated in CHM and sperm and methylated in OT (**FIG. 3A**). A BLAST search showed that the CpG island or GC rich region spanned the putative promoter region and body of a gene predicted by GENSCAN (http://genes.mit.edu/GENSCAN), and included 1638 nucleotides encoding 546 amino acids (**FIG. 6**). BLAST searches of GenBank and Celera databases using the predicted sequences revealed that the predicted gene showed 43% amino acid identity to human transcription elongation factor B (SIII) polypeptide 3 (TCEB3), also known as Elongin A. The novel sequence was even more closely related to a previously identified homolog of Elongin A termed Elongin A2, or TCEB3L, showing 79% amino acid sequence identity to human transcription elongation factor (SIII) Elongin A2 (TCEB3L). We therefore term this gene Elongin A3. An alternative term is TCEB3L2, but for this term to apply, the nomenclature committee will need to rename TCEB3L (Elongin A2) TCEB3L1.

[0084] Analysis of allele-specific expression showed monoallelic expression in lung, brain, placenta, and spinal cord, with preferential expression from the maternal allele (FIGS. 7A-D). There was incomplete preferential expression from the maternal allele in two of three kidneys (**FIGS. 7A, C**), and absence of imprint-specific gene expression in one kidney and in the intestine or liver (**FIGS. 7B, C**, D). Thus, Elongin A3 shows tissue-specific imprinting, at least in prenatal development.

[0085] In another aspect, the present invention provides an isolated polynucleotide that includes a nucleotide sequence unmethylated in nucleic acid of paternal origin and methylated in nucleic acid of maternal origin. The polynucleotiode is about 1638 nucleotides encoding about 546 amino acids and has about 79% amino acid sequence identity to Elongin A2. In a embodiment, the polynucleotide is set forth in SEQ ID NO: 1. This polynucleotide appears to be polymorphic at position 910, which can be G or A. In another embodiment, the polynucleotide encodes a polypeptide as set forth in SEQ ID NO: 2.

[0086] The polynucleotide of SEQ ID NO: 1 is a novel imprinted gene that was identified using methods of the present invention (as illustrated in the Examples below). The CpG island or GC rich region gDMR 2-78 was localized to 18q21 (**FIG. 5**) and was completely methylated in all somatic fetal and adult tissues tested (**FIG. 2**). However, this CpG island or GC rich region was unmethylated in complete hydatidiform moles (CHM) and sperm and methylated in ovarian teratomas (OT) (**FIG. 3A**). A BLAST search showed that the CpG island or GC rich region spanned the putative promoter region and body of a gene predicted by GENSCAN (http://genes.mit.edu/GENSCAN), and included 1638 nucleotides encoding 546 amino acids (**FIG. 6**). BLAST searches of GenBank and Celera databases using the predicted sequences revealed that the predicted gene showed 43% amino acid identity to human transcription elongation factor B (SIII) polypeptide 3 (TCEB3), also known as Elongin A. The novel sequence was even more closely related to a previously identified homolog of Elongin A termed Elongin A2, or TCEB3L, showing 79% amino acid sequence identity to human transcription elongation factor (SIII) Elongin A2 (TCEB3L). We therefore refer to this gene herein as Elongin A3, alternatively TCEB3L2.

[0087] The Elongin A3 gene exhibits monoallelic expression in lung, brain, placenta, spinal cord, and some kidneys, with preferential expression from the maternal allele (FIGS. 7A-D). The gene also exhibits an absence of imprint-specific gene expression in one kidney and in the intestine or liver (**FIGS. 7B, C**, D). Thus, Elongin A3 shows tissue-specific imprinting, at least in prenatal development. Based on this expression pattern, the Elongin A3 gene is useful for example, as a marker for tissue-specific imprinting.

[0088] It is known from previous studies that the elongin (SIII) complex, which includes elonging A1, strongly stimulates the rate of elongation by RNA polymerase II by suppressing transient pausing by polymerase at many sites along the DNA. Elongin (SIII) is composed of a transcriptionally active A subunit and two small regulatory B and C subunits, which bind stably to each other to form a binary complex that interacts with elongin A and strongly induces its transcriptional activity. Elongin A1, B, and, C are highly conserved between mammals and yeast (Aso, T. et al., Biochem Biophys Res Commun., 241(2):334-40 (1997)).

[0089] The elongin (SIII) complex is known to be a potential target for negative regulation by the von Hippel-Lindau (VHL) tumor suppressor protein, which is capable of binding stably to the elongin BC complex and preventing it from activating elongin A. Additionally, it is known that both the elongin A elongation activation domain and the VHL tumor suppressor protein interact with the elongin BC complex through a conserved elongin BC binding site motif that is essential for induction of elongin A activity by elongin BC and for tumor suppression by the VHL protein (Aso T., et al., EMBO J, 15(20):5557-66 (1996)). Elongin A2 is also known to stimulate transcription by RNA Polymerase II.

[0090] Based on these results with elongin Al and elongin A2, elongin A3 polynucleotides and polypeptides of the present invention have utility in in vitro transcription reactions, in stimulating transcription by RNA polymerase. Additionally, elongin A3 polynucleotides and polypeptides of the invention have utility in identifying additional tumor suppressor genes which interact with transcriptional machinery since it is known that at least one transcription factor interacts with an elongin complex, as discussed above.

[0091] As used herein, the term "isolated,""substantially purified" or "substantially pure" means that the molecule being referred to, for example, a polyppeptide or a polynucleotide, is in a form that is relatively free of proteins, nucleic acids, lipids, carbohydrates or other materials with which it is naturally associated. Generally, a substantially pure polypeptide, polynucleotide, or other molecule constitutes at least twenty percent of a sample, generally constitutes at least about fifty percent of a sample, usually constitutes at least about eighty percent of a sample, and particularly constitutes about ninety percent or ninety-five

percent or more of a sample. A determination that a polypeptide or a polynucleotide of the invention is substantially pure can be made using well known methods, for example, by performing electrophoresis and identifying the particular molecule as a relatively discrete band. A substantially pure polynucleotide, for example, can be obtained by cloning the polynucleotide, or by chemical or enzymatic synthesis. A substantially pure polypeptide can be obtained, for example, by using methods of protein purification, such as chromatographic or electrophoretic methods.

[0092] In another aspect, the present invention provides an isolated polypeptide according to SEQ ID NO:2.

[0093] In another aspect, the invention provides, an isolated or purified, polynucleotide that encodes an elongin A3 polypeptide described herein (SEQ ID NO: 2), e.g., a full-length protein or a fragment thereof, e.g., a biologically active portion of the elongin A3 protein. Also included is a nucleic acid fragment suitable for use as a hybridization probe, which can be used, e.g., to identify a nucleic acid molecule encoding a polypeptide of the invention, an elongin A3 mRNA, and fragments suitable for use as primers, e.g., PCR primers for the amplification or mutation of nucleic acid molecules. In one embodiment, an isolated polynucleotide of the invention includes the nucleotide sequence shown in SEQ ID NO: 1, or a portion of any of these nucleotide sequences.

[0094] In another embodiment, an isolated polynucleotide of the invention includes a nucleic acid molecule which is a complement of the nucleotide sequence shown in SEQ ID NO: 1, or a portion of any of these nucleotide sequences. In other embodiments, the nucleic acid molecule of the invention is sufficiently complementary to the nucleotide sequence shown in SEQ ID NO: 1, such that it can hybridize (e.g., under high stringency conditions) to the nucleotide sequence shown in SEQ ID NO: 1 or 3, thereby forming a stable duplex.

[0095] In one embodiment, an isolated polynucleotide of the present invention includes a nucleotide sequence which is at least about: 60%, 65%, 70%, 75%, 80%, 85%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, 99%, or more homologous to the entire length of the nucleotide sequence shown in SEQ ID NO: 1, or a portion, preferably of the same length, of any of these nucleotide sequences.

[0096] A nucleic acid molecule of the invention can include only a portion of the nucleic acid sequence of SEQ ID NO: 1. For example, such a nucleic acid molecule can include a fragment which can be used as a probe or primer or a fragment encoding a portion of an elongin A3 protein, e.g., an immunogenic or biologically active portion of an elongin A3 protein. The nucleotide sequence determined from the cloning of the elongin A3 gene allows for the generation of probes and primers designed for use in identifying and/or cloning other elongin A3 family members, or fragments thereof, as well as elongin A3 homologues, or fragments thereof, from other species.

[0097] In another embodiment, a nucleic acid encodes an polypeptide fragment of elongin A3 described herein. Nucleic acid fragments can encode a specific domain or site described herein or fragments thereof, particularly fragments thereof which are at least 100, 150, 200, 210, 220, 230, 240, 250, 300, 350, 400, 450, 500, 550, or 546 amino acids in length. Nucleic acid fragments should not to be construed as encompassing those fragments that may have been disclosed prior to the invention.

[0098] A nucleic acid fragment can include a sequence corresponding to a domain, region, or functional site described herein. A nucleic acid fragment can also include one or more domain, region, or functional site described herein. Thus, for example, an elongin A3 nucleic acid fragment can include a sequence corresponding to a domain that binds other elongins, similar to the domain of elongin A1 which binds elongins B and C. Elongin A3 probes and primers are provided. Typically a probe/primer is an isolated or purified oligonucleotide. The oligonucleotide typically includes a region of nucleotide sequence that hybridizes under moderate, or preferably high stringency conditions to at least about 7, 12 or 15, preferably about 20 or 25, more preferably about 30, 35, 40, 45, 50, 55, 60, 65, or 75 consecutive nucleotides of a sense or antisense sequence of SEQ ID NO: 1, or of a naturally occurring allelic variant or mutant of SEQ ID NO: 1.

[0099] In a embodiment the nucleic acid is a probe which is at least 5 or 10, and less than 200, more preferably less than 100, or less than 50, base pairs in length. It should be identical, or differ by 1, or less than in 5 or 10 bases, from a sequence disclosed herein. If alignment is needed for this comparison the sequences should be aligned for maximum homology. "Looped" out sequences from deletions or insertions, or mismatches, are considered differences.

[0100] In another embodiment a set of primers is provided, e.g., primers suitable for use in a PCR, which can be used to amplify a selected region of a elongin A3 sequence, e.g., a domain, region, site or other sequence described herein. The primers should be at least 5, 10, or 50 base pairs in length and less than 100, or less than 200, base pairs in length. The primers should be identical, or differs by one base from a sequence disclosed herein or from a naturally occurring variant.

[0101] A nucleic acid fragment can encode an epitope bearing region of a polypeptide described herein. A nucleic acid fragment encoding a "biologically active portion of a elongin A3 polypeptide" can be prepared by isolating a portion of the nucleotide sequence of SEQ ID NO: 1, which encodes a polypeptide having a elongin A3 biological activity (e.g., the biological activities of the elongin A3 protein is described herein), expressing the encoded portion of the elongin A3 protein (e.g., by recombinant expression in vitro) and assessing the activity of the encoded portion of the elongin A3 protein. For example, a nucleic acid fragment encoding a biologically active portion of elongin A3 includes a domain that binds with other elongs such as elongin B or elongin C. A nucleic acid fragment encoding a biologically active portion of a elongin A3 polypeptide, can comprise a nucleotide sequence which is greater than 300 or more nucleotides in length.

[0102] In certain embodiments, a nucleic acid of the invention includes a nucleotide sequence which is about 300, 400, 500, 600, 700, 800, 900, 1000, 1200, 1400, 1600, 1800, 1900, or more nucleotides in length and hybridizes under moderate or high stringency conditions to a nucleic acid molecule of SEQ ID NO: 1.

[0103] In embodiments, the fragment includes at least one, and preferably at least 5, 10, 15, 25, 50, 75, 100, 200, 300,

500, 1000, 1500, or 1620 nucleotides encoding a protein including 5, 10, 15, 20, 25, 30, 40, 50, 100, 200, 210, 220, 230, 240, 250, 300, 350, 400, 450, 500, or 546 consecutive amino acids of SEQ ID NO: 2.

[0104] The invention further encompasses nucleic acid molecules that differ from the nucleotide sequence shown in SEQ ID NO: 1. Such differences can be due to degeneracy of the genetic code, and result in a nucleic acid which encodes the same elongin A3 protein as that encoded by the nucleotide sequence disclosed herein. In another embodiment, an isolated polynucleotide of the invention has a nucleotide sequence encoding a protein having an amino acid sequence which differs, by at least 1, but less than 5, 10, 20, 50, or 100 amino acid residues that shown in SEQ ID NO: 2. If alignment is needed for this comparison the sequences should be aligned for maximum homology. "Looped" out sequences from deletions or insertions, or mismatches, are considered differences.

[0105] As used herein, the term "selective hybridization" or "selectively hybridize" refers to hybridization under moderately stringent or highly stringent physiological conditions, which can distinguish related nucleotide sequences from unrelated nucleotide sequences.

[0106] As known in the art, in nucleic acid hybridization reactions, the conditions used to achieve a particular level of stringency will vary, depending on the nature of the nucleic acids being hybridized. For example, the length, degree of complementarity, nucleotide sequence composition (for example, relative GC:AT content), and nucleic acid type, i.e., whether the oligonucleotide or the target nucleic acid sequence is DNA or RNA, can be considered in selecting hybridization conditions. An additional consideration is whether one of the nucleic acids is immobilized, for example, on a filter. Methods for selecting appropriate stringency conditions can be determined empirically or estimated using various formulas, and are well known in the art (see, for example, Sambrook et al., supra, 1989).

[0107] An example of progressively higher stringency conditions is as follows: 2×SSC/0.1% SDS at about room temperature (hybridization conditions); 0.2×SSC/0.1% SDS at about room temperature (low stringency conditions); 0.2×SSC/0.1% SDS at about 42° C. (moderate stringency conditions); and 0.1×SSC at about 68° C. (high stringency conditions). Washing can be carried out using only one of these conditions, for example, high stringency conditions, or each of the conditions can be used, for example, for 10 to 15 minutes each, in the order listed above, repeating any or all of the steps listed.

[0108] Nucleic acids of the invention can be chosen for having codons, which are compatible, or noncompatible, for a particular expression system, e.g., the nucleic acid can be one in which at least one codon, or at least 10%, or 20% of the codons has been altered such that the sequence is optimized for expression in *E. coli*, yeast, human, insect, or CHO cells, for example.

[0109] Nucleic acid variants can be naturally occurring, such as allelic variants (same locus), homologs (different locus), and orthologs (different organism) or can be non naturally occurring. Non-naturally occurring variants can be made by mutagenesis techniques, including those applied to polynucleotides, cells, or organisms. The variants can contain nucleotide substitutions, deletions, inversions and insertions. Variation can occur in either or both the coding and non-coding regions. The variations can produce both conservative and non-conservative amino acid substitutions (as compared in the encoded product).

[0110] Orthologs, homologs, and allelic variants can be identified using methods known in the art. These variants comprise a nucleotide sequence encoding a polypeptide that is 50%, at least about 55%, typically at least about 70-75%, more typically at least about 80-85%, and most typically at least about 90-95% or more identical to the nucleotide sequence shown in SEQ ID NO: 2 or a fragment of this sequence. Such nucleic acid molecules can readily be identified as being able to hybridize under moderate, or preferably high stringency condition, to the nucleotide sequence shown in SEQ ID NO: 2 or a fragment of the sequence. Nucleic acid molecules corresponding to orthologs, homologs, and allelic variants of the elongin A3 cDNAs of the invention can further be isolated by mapping to the same chromosome or locus as the elongin A3 gene.

[0111] Allelic variants of elongin A3, e.g., human elongin A3, include both functional and non-functional proteins. Functional allelic variants are naturally occurring amino acid sequence variants of the elongin A3 protein within a population that maintain the ability to increase the speed of transcription by RNA polymerase. Functional allelic variants will typically contain only conservative substitution of one or more amino acids of SEQ ID NO: 2, or substitution, deletion or insertion of non-critical residues in non-critical regions of the protein. Non-functional allelic variants are naturally-occurring amino acid sequence variants of the elongin A3, e.g., human elongin A3, protein within a population that do not have the ability to participate in redox reactions or molecular chaperone interactions. Non-functional allelic variants will typically contain a non-conservative substitution, a deletion, or insertion, or premature truncation of the amino acid sequence of SEQ ID NO: 2, or a substitution, insertion, or deletion in critical residues or critical regions of the protein. As disclosed hererin, a polymorphism identified for elongin A3 includes a G or A residue at position 910.

[0112] In another aspect, the invention features, an isolated polynucleotide which is antisense to elongin A3. An "antisense" nucleic acid can include a nucleotide sequence which is complementary to a "sense" nucleic acid encoding a protein, e.g., complementary to the coding strand of a double-stranded cDNA molecule or complementary to an mRNA sequence. The antisense nucleic acid can be complementary to an entire elongin A3 coding strand, or to only a portion thereof.

[0113] An antisense nucleic acid can be designed such that it is complementary to the entire coding region of elongin A3 mRNA, but more preferably is an oligonucleotide which is antisense to only a portion of the coding or noncoding region of elongin A3 mRNA. For example, the antisense oligonucleotide can be complementary to the region surrounding the translation start site of elongin A3 mRNA, e.g., between the −10 and +10 regions of the target gene nucleotide sequence of interest. An antisense oligonucleotide can be, for example, about 7, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, or more nucleotides in length.

[0114] The invention also provides detectably labeled oligonucleotide primer and probe molecules. Typically, such

labels are chemiluminescent, fluorescent, radioactive, or calorimetric. The invention also includes molecular beacon oligonucleotide primer and probe molecules having at least one region which is complementary to a elongin A3 nucleic acid of the invention, two complementary regions one having a fluorophore and one a quencher such that the molecular beacon is useful for quantitating the presence of the elongin A3 nucleic acid of the invention in a sample. Molecular beacon nucleic acids are described, for example, in Lizardi et al., U.S. Pat. No. 5,854,033; Nazarenko et al., U.S. Pat. No. 5,866,336, and Livak et al., U.S. Pat. No. 5,876,930.

[0115] In another aspect, the invention features, an isolated elongin A3 protein, or fragment, e.g., a biologically active portion, for use as immunogens or antigens to raise or test (or more generally to bind) anti-elongin A3 antibodies. Elongin A3 protein can be isolated from cells or tissue sources using standard protein purification techniques. Elongin A3 protein or fragments thereof can be produced by recombinant DNA techniques or synthesized chemically.

[0116] Polypeptides of the invention include those which arise as a result of the existence of multiple genes, alternative transcription events, alternative RNA splicing events, and alternative translational and post-translational events. The polypeptide can be expressed in systems, e.g., cultured cells, which result in substantially the same post-translational modifications present when expressed the polypeptide is expressed in a native cell, or in systems which result in the alteration or omission of post-translational modifications, e.g., glycosylation or cleavage, present when expressed in a native cell.

[0117] In an embodiment, a elongin A3 polypeptide has a molecular weight, e.g., a deduced molecular weight, preferably ignoring any contribution of post translational modifications, amino acid composition or other physical characteristic of SEQ ID NO: 2 or it has an overall sequence similarity of at least 50%, preferably at least 60%, more preferably at least 70, 80, 90, or 95%, with a polypeptide a of SEQ ID NO: 2.

[0118] In another aspect, the invention provides an anti-elongin A3 antibody, or a fragment thereof (e.g., an antigen-binding fragment thereof). The term "antibody" as used herein refers to an immunoglobulin molecule or immunologically active portion thereof, i.e., an antigen-binding portion. As used herein, the term "antibody" refers to a protein comprising at least one, and preferably two, heavy (H) chain variable regions (abbreviated herein as VH), and at least one and preferably two light (L) chain variable regions (abbreviated herein as VL). The anti-elongin A3 antibody can be a polyclonal or a monoclonal antibody. In other embodiments, the antibody can be recombinantly produced, e.g., produced by phage display or by combinatorial methods.

[0119] In embodiments an antibody can be made by immunizing with purified elongin A3 antigen, or a fragment thereof, e.g., a fragment described herein, membrane associated antigen, tissue, e.g., crude tissue preparations, whole cells, preferably living cells, lysed cells, or cell fractions, e.g., membrane fractions. A full-length elongin A3 protein or, antigenic peptide fragment of elongin A3 can be used as an immunogen or can be used to identify anti-elongin A3 antibodies made with other immunogens, e.g., cells, membrane preparations, and the like. The antigenic peptide of elongin A3 should include at least 8 amino acid residues of the amino acid sequence shown in SEQ ID NO: 2 and encompasses an epitope of elongin A3. Preferably, the antigenic peptide includes at least 10 amino acid residues, more preferably at least 15 amino acid residues, even more preferably at least 20 amino acid residues, and most preferably at least 30 amino acid residues.

[0120] An anti-elongin A3 antibody (e.g., monoclonal antibody) can be used to isolate elongin A3 by standard techniques, such as affinity chromatography or immunoprecipitation. Moreover, an anti-elongin A3 antibody can be used to detect elongin A3 protein (e.g., in a cellular lysate or cell supernatant) in order to evaluate the abundance and pattern of expression of the protein. Anti-elongin A3 antibodies can be used diagnostically to monitor protein levels in tissue as part of a clinical testing procedure, e.g., to determine the efficacy of a given treatment regimen.

[0121] The invention also includes cell lines, e.g., hybridomas, which make an anti-elongin A3 antibody, e.g., and antibody described herein, and method of using said cells to make a elongin A3 antibody.

[0122] In another aspect, the invention includes, vectors, preferably expression vectors, containing a nucleic acid encoding an elongin A3 polypeptide, preferably the elongin A3 polypeptide of SEQ ID NO: 2. The vector can be capable of autonomous replication or it can integrate into a host DNA. Viral vectors include, e.g., replication defective retroviruses, adenoviruses and adeno-associated viruses.

[0123] A vector can include a elongin A3 nucleic acid in a form suitable for expression of the nucleic acid in a host cell. Preferably the recombinant expression vector includes one or more regulatory sequences operatively linked to the nucleic acid sequence to be expressed. The term "regulatory sequence" includes promoters, enhancers and other expression control elements (e.g., polyadenylation signals). Regulatory sequences include those which direct constitutive expression of a nucleotide sequence, as well as tissue-specific regulatory and/or inducible sequences. The design of the expression vector can depend on such factors as the choice of the host cell to be transformed, the level of expression of protein desired, and the like. The expression vectors of the invention can be introduced into host cells to thereby produce proteins or polypeptides, including fusion proteins or polypeptides, encoded by nucleic acids as described herein (e.g., elongin A3 proteins, mutant forms of elongin A3 proteins, fusion proteins, and the like).

[0124] The recombinant expression vectors of the invention can be designed for expression of elongin A3 proteins in prokaryotic or eukaryotic cells. For example, polypeptides of the invention can be expressed in *E. coli*, insect cells (e.g., using baculovirus expression vectors), yeast cells or mammalian cells. Suitable host cells are discussed further in Goeddel, (1990) Gene Expression Technology: Methods in Enzymology 185, Academic Press, San Diego, Calif. Alternatively, the recombinant expression vector can be transcribed and translated in vitro, for example using T7 promoter regulatory sequences and T7 polymerase.

[0125] Expression of proteins in prokaryotes is most often carried out in *E. coli* with vectors containing constitutive or inducible promoters directing the expression of either fusion

or non-fusion proteins. Fusion vectors add a number of amino acids to a protein encoded therein, usually to the amino terminus of the recombinant protein. Such fusion vectors typically serve three purposes: 1) to increase expression of recombinant protein; 2) to increase the solubility of the recombinant protein; and 3) to aid in the purification of the recombinant protein by acting as a ligand in affinity purification. Often, a proteolytic cleavage site is introduced at the junction of the fusion moiety and the recombinant protein to enable separation of the recombinant protein from the fusion moiety subsequent to purification of the fusion protein. Such enzymes, and their cognate recognition sequences, include Factor Xa, thrombin and enterokinase. Typical fusion expression vectors include pGEX (Pharmacia Biotech Inc; Smith, D. B. and Johnson, K. S. (1988) Gene 67:31-40), pMAL (New England iolabs, Beverly, Mass.) and pRIT5 (Pharmacia, Piscataway, N.J.) which fuse glutathione S-transferase (GST), maltose E binding protein, or protein A, respectively, to the target recombinant protein.

[0126] The elongin A3 expression vector can be a yeast expression vector, a vector for expression in insect cells, e.g., a baculovirus expression vector or a vector suitable for expression in mammalian cells. When used in mammalian cells, the expression vector's control functions can be provided by viral regulatory elements. For example, commonly used promoters are derived from polyoma, Adenovirus 2, cytomegalovirus and Simian Virus 40.

[0127] Another aspect the invention provides a host cell which includes a nucleic acid molecule described herein, e.g., a elongin A3 nucleic acid molecule within a recombinant expression vector or a elongin A3 nucleic acid molecule containing sequences which allow it to homologously recombine into a specific site of the host cell's genome. The terms "host cell" and "recombinant host cell" are used interchangeably herein. Such terms refer not only to the particular subject cell but to the progeny or potential progeny of such a cell. Because certain modifications may occur in succeeding generations due to either mutation or environmental influences, such progeny may not, in fact, be identical to the parent cell, but are still included within the scope of the term as used herein.

[0128] A host cell can be any prokaryotic or eukaryotic cell. For example, a elongin A3 protein can be expressed in bacterial cells (such as *E. coli*), insect cells, yeast or mammalian cells (such as Chinese hamster ovary cells (CHO) or COS cells (African green monkey kidney cells CV-1 origin SV40 cells; Gluzman (1981) CellI23:175-182)). Other suitable host cells are known to those skilled in the art.

[0129] Vector DNA can be introduced into host cells via conventional transformation or transfection techniques. As used herein, the terms "transformation" and "transfection" are intended to refer to a variety of art-recognized techniques for introducing foreign nucleic acid (e.g., DNA) into a host cell, including calcium phosphate or calcium chloride co-precipitation, DEAE-dextran-mediated transfection, lipofection, or electroporation.

[0130] A host cell of the invention can be used to produce (i.e., express) an elongin A3 protein. Accordingly, the invention further provides methods for producing an elongin A3 protein using the host cells of the invention. In one embodiment, the method includes culturing the host cell of the invention (into which a recombinant expression vector encoding a elongin A3 protein has been introduced) in a suitable medium such that an elongin A3 protein is produced. In another embodiment, the method further includes isolating an elongin A3 protein from the medium or the host cell.

[0131] In another aspect, the invention features, a cell or purified preparation of cells which include a elongin A3 transgene, or which otherwise misexpress elongin A3. The cell preparation can consist of human or non-human cells, e.g., rodent cells, e.g., mouse or rat cells, rabbit cells, or pig cells. In embodiments, the cell or cells include a elongin A3 transgene, e.g., a heterologous form of a elongin A3, e.g., a gene derived from humans (in the case of a non-human cell).

[0132] The invention provides non-human transgenic animals. Such animals are useful for studying the function and/or activity of a elongin A3 protein and for identifying and/or evaluating modulators of elongin A3 activity. As used herein, a "transgenic animal" is a non-human animal, preferably a mammal, more preferably a rodent such as a rat or mouse, in which one or more of the cells of the animal includes a transgene. Other examples of transgenic animals include non-human primates, sheep, dogs, cows, goats, chickens, amphibians, and the like. A transgene is exogenous DNA or a rearrangement, e.g., a deletion of endogenous chromosomal DNA, which preferably is integrated into or occurs in the genome of the cells of a transgenic animal. A transgene can direct the expression of an encoded gene product in one or more cell types or tissues of the transgenic animal, other transgenes, e.g., a knockout, reduce expression. Thus, a transgenic animal can be one in which an endogenous elongin A3 gene has been altered by, e.g., by homologous recombination between the endogenous gene and an exogenous DNA molecule introduced into a cell of the animal, e.g., an embryonic cell of the animal, prior to development of the animal.

[0133] Elongin A3 proteins or polypeptides can be expressed in transgenic animals or plants, e.g., a nucleic acid encoding the protein or polypeptide can be introduced into the genome of an animal. In embodiments the nucleic acid is placed under the control of a tissue specific promoter, e.g., a milk or egg specific promoter, and recovered from the milk or eggs produced by the animal. Suitable animals are mice, pigs, cows, goats, and sheep.

[0134] The invention also provides a method for identifying the DNA methylation status at a cytosine residue of a CpG sequence in a genomic DNA sample from the subject, wherein hypomethylation of CpG sequences compared to a methylated DNA control sample is indicative of a disease present within the subject. In one aspect, the method is characterized in that a set of CpG positions comprises at least 3 CpG positions that are located in the regulatory region of the same gene. In another aspect, the method is characterized in that the methylation state of at least 3 different sets of CpG positions is identified.

[0135] The following examples are intended to illustrate but not limit the invention.

EXAMPLE 1

[0136] Isolation of Normally Methylated CpG Islands or GC Rich Regions from a Genome-Wide Screen

[0137] This example illustrates a method for isolating normally methylated CpG islands or GC rich regions in a genome-wide screen.

[0138] Isolation and Identification of Methylated CpG Islands or GC Rich Regions from Genomic DNA

[0139] A two-step cloning procedure was used for isolating and identifying methylated CpG islands or GC rich regions from genomic DNA. In the first step, 200 μg of genomic DNA were digested overnight with 1000 units of Hpa II (LTI) followed by a 5-h digest with 600 units of Mse I (NEB), according to the manufacturer's conditions, and the volume was reduced using a SpeedVac concentrator (Savant). Fragments 1 kb were size selected using a Chromaspin+TE, 400 column (Clontech), and fragments between 1-9 kb were purified from a 0.8% gel by electroelution and an Elutip-D column (S&S). The eluate was ethanol precipitated, cloned into the compatible Nde I site of pGEM-4Z, which was first modified to abolish the Sma I site, transformed into the competent cells of the restriction-deficient strain XL2-Blue MRF (Stratagene), and plated onto LB-ampicillin agar plates. Library DNA was prepared directly from plates using a plasmid Maxi kit (Qiagen). In the second step, 100 μg of the Mse I library DNA were digested with 1,000 U of Eag I (NEB) according to the manufacturer's conditions. The digest was ethanol precipitated, and 100-1500-bp fragments were size-selected by purification from a 1.5% agarose gel, cloned into the Eag I site of pBC (Stratagene), and transformed into XL1 -Blue MRF' (Stratagene).

[0140] DNA Sequencing

[0141] DNA sequencing was performed using an ABI 377 automated sequencer following protocols recommended by the manufacturer (Perkin-Elmer). The sequences were analyzed by BLAST search (Altschul et al. 1990) of the GenBank and Celera databases.

[0142] Southern Hybridization

[0143] Genomic DNA was digested with Mse I alone or Mse I together with a methylcytosine-sensitive (Hpa II, LTI, or Sma I, NEB) or methyl-insensitive (Msp I or Xma I, NEB) restriction endonuclease according to the manufacturer's conditions. Southern hybridization was performed as described (Dyson 1991).

[0144] Imprinting Analysis of Elongin A3 Gene

[0145] Fetal tissues and matched maternal decidua were obtained from the University of Washington Fetal Tissue Bank. We identified polymorphisms by sequencing fetal and maternal PCR amplified genomic DNA. The following conditions were used for PCR amplifications: 95° C., 2 min; then 40 cycles of 95° C. 1 min, 60° C. 30 sec, 72° C. min; then 72° C. for 9 min. Total RNA was isolated from fetal tissues using RNeasy mini kit (Qiagen). To eliminate DNA contamination from RNA preparations, samples were treated with preamplification-grade DNase I (Invitrogen) according to supplied protocols. RT-PCR was carried out using the Superscript II preamplification system (Invitrogen) and was performed for each sample in the presence and absence (negative controls) of RT. cDNA samples were sequenced only when no bands were obtained with the negative controls. The primers used for the imprinting analysis were EL2AL-1093-1112F: 5'-TCT GCT GTC CGC TTT TGA GG-3' (SEQ ID NO:32) and EL2AL-1526-1550R: 5'-ATC GGA TTT TCG TGG TCA CTA CTC G-3' (SEQ ID NO:33). DNA and cDNA sequencing was run on an ABI-377 automated sequencer following protocols recommended by the manufacturer (Perkin-Elmer).

[0146] Isolation of Normally Methylated CpG Islands or GC Rich Regions

[0147] A restriction enzyme-based strategy was chosen for isolating methylated CpG islands over a PCR-based strategy, to avoid known problems of amplification bias against GC-rich sequences, and to obtain larger clone inserts than would be possible by a PCR-based approach. DNA from tissue from a male was used, to avoid cloning methylated CpG islands from the inactive X chromosome, and to avoid cell culture-induced DNA methylation. The tissue chosen was a Wilms tumor, because this approach would identify either normally methylated CpG islands or those methylated specifically in this tumor, which is of interest to our laboratory. The plan was to determine after cloning these sequences whether they were methylated in normal cells or in tumors. The first step of the approach (**FIG. 1**) involved double digestion with Mse I, which recognizes the sequence TTAA and Hpa II, which recognizes the sequence CCGG at unmethylated sites. Mse I digests DNA between CpG islands, and Hpa II digests unmethylated CpG islands into small fragments, as it has a 4-bp recognition sequence. These digestions were followed by gel purification of fragments >1 kb in length. These initial digestions and purification were predicted by computer analysis of GenBank to enrich ~10-fold for CpG islands, and enrichment of known methylated CpG islands (near imprinted genes) was confirmed by Southern blot hybridization. At the same time, this step eliminates all unmethylated CpG islands because of the methylcytosine sensitivity of Hpa II. The restriction fragments obtained by this first step then were cloned into the restriction-negative strain XL2-Blue MRF' to avoid bacterial digestion of methylated genomic DNA, and the resulting genomic library was termed the "Mse library." The second cloning step (**FIG. 1**) involved further enrichment of CpG islands by digesting the purified Mse I library DNA with an infrequently cutting restriction endonucleases (i.e., recognizing 6 bp CG-rich sequences) specific for sequences common to CpG islands, to isolate relatively large fragments of CpG islands that are normally methylated (i.e., survived the first cloning step), but are now unmethylated in the Mse library and therefore amenable to digestion and subcloning. Most of the work described here was performed by using Eag I (recognition sequence CGGCCG) in this second step, and subcloning Eag I fragments in three size classes separated by agarose gel electrophoresis (100-500 bp, 500-1000 bp, >1000 bp), and the resulting library was termed the Eag library.

[0148] Methylated CpG Islands Within Interspersed Repeats

[0149] The primary goal was to identify unique methylated CpG islands throughout the genome. However, it quickly became apparent that most of the clones in the Eag library represented high copy number methylated CpG islands. The majority of these clones were derived from a sequence termed SVA, which constituted 70% of the Eag I library, and was not previously known to be methylated. The little-known SVA retroposon contains a GC-rich VNTR region, which embodies a CpG rich region between an Alu-derived region and an LTR-derived region. Only three such elements had previously been described (Kawajiri et al. 1986; Zhu et al. 1992; Shen 1994), although their methylation has not been characterized. A probe termed SVA-U was designed, which was unique to the SVA and present in all of

the SVA elements, to analyze copy number and methylation of this sequence in genomic DNA. The copy number was estimated by quantitative Southern hybridization to be 5000 per haploid genome. The SVA elements were found to be completely methylated in all adult somatic tissues examined, including peripheral blood lymphocytes, kidney, adrenal, liver and lung. A somewhat less abundant high copy repeat, representing an additional 20% of the Eag I library, corresponded to the nontranscribed intergenic spacer of ribosomal DNA, which was a known methylated repetitive sequence (Brock and Bird 1997), suggesting that ribosomal gene methylation may be more extensive than was previously suspected. The focus of the current study was on the unique methylated CpG islands that were identified after excluding these sequences.

EXAMPLE 2

[0150] Methylation Analysis of Novel Single Copy CpG Islands or GC Rich Regions

[0151] This example illustrates that the methods of the present invention for identifying and isolating methylated CpG islands or GC rich regions are effective for identifying imprinted genes.

[0152] Isolation and identification of methylated CpG islands from genomic DNA was performed as described in Example 1, except that to eliminate methylated CpG islands that corresponded to dispersed repetitive sequences, the Mse I library was derived by adding restriction enzymes designed to cleave those sequences and render them unclonable. For 28S and ribosomal DNA, we used Asc I. For SVA, we used Dra III+Sal I, followed by either Acc I or TthIII1.

[0153] To isolate single-copy clones, we re-derived the Mse library, adding restriction endonucleases designed to cleave repeat sequences described above, rendering them unclonable (see Methods). After eliminating redundant clones, 62 unique clones were characterized. All of the sequences were GC-rich, i.e., with a measured (C+G)/N>50%, and they ranged in GC content from 55 to 79%. Forty-three (69%) of the clones showed an observed to expected CpG ratio >0.6, meeting the formal definitional requirement of a CpG rich region, and they were characterized further. Nevertheless, most of the remaining clones showed an observed to expected CpG ratio >0.5.

[0154] As the original source of DNA was a Wilm's tumor, we had no a priori knowledge of the methylation status of these sequences in normal tissue. Surprisingly, all of the sequences were methylated in normal lymphocyte DNA **(FIG. 2A)**. Methylation was not restricted to lymphocyte DNA, as it also was observed in both adult and fetal tissues, including brain, gut, kidney, liver, lung, and skin **(FIG. 2B)**. Thus, these sequences represented normally methylated CpG islands.

[0155] To determine whether the CpG islands were differentially methylated in the maternal and paternal germline, 30 of the clones were individually hybridized to Southern blots of DNA isolated from ovarian teratomas (OT) and complete hydatidiform moles (CHM), which are of uniparental maternal and paternal origin, respectively (CHM DNA was exhausted at that point). Thirteen clones exhibited methylation in the OT but not or significantly less so in the CHM (Table 1). For example, CpG rich region 2-78 showed complete digestion after Hpa II treatment of genomic DNA isolated from sperm and CHM, similar to the pattern after Msp I digestion **(FIG. 3A)**. In contrast, 2-78 showed an identical pattern after Mse I+Hpa II digestion, as after Mse I alone, in OT **(FIG. 3A)**. Similarly, **FIG. 3A** shows OT-specific methylation of CpG islands 3-30, 1-13, 4-6, and 2-48, with relative lack of methylation in CHM. These sequences therefore represent differentially methylated regions, because of their different pattern of methylation in germline tissues of male (sperm and CHM) and female (OT) origin. Because many of these sequences also are methylated in somatic tissues, we refer to them as gDMR's (germline differentially methylated regions). All of the gDMR sequences were methylated in OT and not CHM. As a negative control, a CpG rich region associated with the RB gene (retinoblastoma) is unmethylated in both CHM and OT. As a positive control, a CpG rich region upstream of the imprinted gene H19 is preferentially methylated in CHM, and a CpG rich region within the imprinted SNRPN gene is methylated in OT **(FIG. 3B)**.

TABLE 1

Methylated CpG Islands Characterized In Detail

Clone ID	SEQ. ID No.	Methylation pattern	Chromosomal location	Associated genes: Accession no.	Associated genes: Predicted function
3-10	3	SMR	1q44	14042537[a]	Similar to Zn finger protein
				14423768	Olfactory receptor 2T1
				hCG1644736	Similar to olfactory receptor 271
				hCG1724357	Similar to olfactory receptor 2T1
3-20	4	SMR	2q36	5174481[a]	Histone deacetylase A
				hCG1651464	
				hCG1656118	
				hCG1651461	
				hCG1651466	
1-19	5	SMR	4p16	37775830[a]	WFS1 (wolframin)
1-41	6	SMR	4q35	hCG1788598[a]	Similar to mouse pair-rule gene ODZ3
				hCG1793025[a]	Hypothetical protein
				hCG1787540	Similar to mouse pair-rule gene ODZ3
4-8	7	SMR	4q35	hCG1788598[a]	Adjacent to but distinct from 1–41
3-4	8	gDMR	6q24	hCG1660630[a]	HYMA1

TABLE 1-continued

Methylated CpG Islands Characterized In Detail

Clone ID	SEQ. ID No.	Methylation pattern	Chromosomal location	Associated genes: Accession no.	Associated genes: Predicted function
4-7	9	SMR	7p22	6806913[a]	Centaurin-α
				hCG1747708	
				hCG1790856	
				hCG1747710	Cytochrome P450 homolog
1-30	10	SMR	7q11.1	hCG1779529[a]	Rab5 exchange factor homolog
				hCG1779527	Antisense to hCG1779529
				hCG1789113	Antisense to hCG1779529
				13642872	60S ribosomal prot L35
				6572672	Putative transcription factor
1-22	11	SMR	7q36	11386149[a]	Tyrosine phosphatase
				hCG1799787	BHLF1 protein
3-2	12	SMR	8p23	hCG1659058[a]	Proline-rich mucin homolog
2-5	13	gDMR	8q21.2	17451956	Similar to antigen GOR
				hCG1757665	
2-48	14	gDMR	9p13	hCG1659616	
1-20	15	gDMR	10q26	13325182a	
				hCG1799063	
3-12	16	SMR	10q26	3122245[a]	Inositol triphosphate phosphatase
				hCG1654478	
3-30	17	gDMR	11q25	17456499[a]	Hypothetical gene
				hCG37607	
				hCG1745526	Hypothetical protein
1-5	18	SMR	13q34	hCG20146[a]	
1-21	19	gDMR	14q32	8393715	Heptacellular cancer candidate
				hCG21408	Similar to Drosophila CLIP-190
2-42	20	gDMR	16p13.1	hCG15669[a]	Ser/Thr protein kinase
3-110	21	SMR	17q25	8400736[a]	βtubulin cofactor D
				10435982	
2-1	22	SMR	17q25	1655842a	Sulfamidase
				14149793	Coiled-coil protein
1-12	23	SMR	17q25	hCG1806389	
				hCG1796817	
2-78	24	gDMR	18q21	13645769[a]	Elongin A3
3-8	25	gDMR	18q21	13645769[a]	Adjacent to but distince from 2–78
2-3	26	SMR	18q23	6912444	Voltage-dependent K+ channel
				1914872	Choline-binding protein
				5326898	RNA Polymerase II CTD phosphatase
1-13	27	gDMR	18q23	hCG1651089[a]	
				6688241	SALL3 Spalt-like zinc finger protein
				hCG20372	ATPase
				1651088	Mucin1 precursor
1-6	28	gDMR	19p13.1	14249150[a]	Hypothetical protein
4-3	29	SMR	19p13.1	9665054[a]	Ser/Thr kinase 11
				hCG23965	Ankyrin repeat protein
				4506715	Ribosomal protein S28
				hCG1794585	
				10732648	Angiopoietin-like protein
1-2	30	SMR	19q13.4	12053197	Zinc finger protein
				4505329	NSF attachment protein
				5901994	Kaptin actin binding protein
				5689511	Na/Ca exchanger protein
				7657128	Glioma tumor suppressor candidate
				7657054	EH-Domain containing protein
				7657130	Glioma tumor suppressor candidate
2-4	31	gDMR	20q12	17484155[a]	Nuclear factor of activated T-cells 2
				hCG1800975	
				hCG1653833	
				13378306	Brain RPTmam4 isoform
				110743	Neurofilament triplet H protein
				7799072	DNA helicase

Accession numbers correspond to GenBank entries, within 10 kb of the CpG island, unless there is no GenBank entry, in which case correspond to Celera entries. One additional SMR could not be mapped.
[a]CpG island lies within the transcript.

[0156] An additional 17 clones identified CpG islands that were methylated equally in OT, CHM, and sperm (Table 1). For example, CpG islands 3-110, 3-10, 2-1, and 1-41 showed an identical pattern after Mse I+Hpa II digestion, as after Mse I alone, in OT and CHM (**FIG. 4**). We termed these sequences SMRs, to connote their comparable methy-

lation in male and female tissue of germline origin. Like the gDMRs, these SMRs were methylated in cells of somatic origin (**FIG. 2A**).

EXAMPLE 3

[**0157**] Chromosomal Location of Methylated CPG Rich Regions and Association with Genes

[**0158**] This Example shows that many SMRs are located near the ends of chromosomes and identifies CpG islands isolated herein that reside near known genes. Chromosomal locations of the identified CpG islands were determined by identifying corresponding Genbank human genomic DNA sequences of known genomic location, using well-known nucleic acid sequence search tools such as BLAST.

[**0159**] The methylated CpG islands identified here were distributed throughout the genome. There was a striking localization of SMRs near the ends of chromosomes. Sixteen of 17 SMRs were localized near the ends of chromosomes, either on the last (n=15) or the penultimate (n=1) subband of the chromosome on which it resided (Table 2). In contrast, of 12 gDMRs that could be mapped (of the 13 gDMRs studied), only four were localized near the ends of chromosomes (Table 2). This difference was highly statistically significant (P=0.0008, Fisher's exact test). The association of SMRs near the ends of chromosomes is consistent with an observation of densely methylated GC-rich sequences near telomeres, although that study did not describe methylated CpG islands (Brock et al. 1999). In addition, there was a segregation of gDMRs and SMRs within compartments of differing genomic composition, i.e., isochores, which are regions of several hundred kilobases of relatively homogeneous GC composition (Bernardi 1995). Approximately 75% of the SMRs fell within high isochore regions (G+C 50%), as might be expected from the high GC content of methylated CpG islands. Surprisingly, however, all of the gDMRs fell within low isochore regions (G+C<50%), i.e., of relatively low GC content, despite the high GC content of the gDMRs themselves (L. Z. Strichman-Almashanu and A. P. Feinberg). This difference was statistically significant (P<0.01, Fisher's exact test). Thus, the gDMRs and SMRs may lie within distinct chromosomal and/or isochore compartments. These results provide the basis for a method to identify epigenetic chromosomal domains. Localization of CpG islands to the telo/subtelo regions, for example, can be used for identifying imprinted gene domains, disease domains (e.g. p16), chromatin regulated genes controlled at a distance, such as telomerase (TERT) or c-myc by CTCF; and developmentally programmed regions essential for organ formation, such as the brain in Lunyak et al. Science. Oct. 24, 2002 for example.

TABLE 2

Band Location of Methylated CpG Rich regions

CpG Rich region	Centromeric	Band Location Midchromosome	Telomeric
GDMR	0	8	4
SMR	1	0	16

[**0160**] There were several examples of nonredundant, unique methylated CpG islands localizing to the same chromosomal region. In two cases, two pairs of sequences were adjacent within the genome. Two SMRs on 4q35, 1-41 and 4-8, were adjacent to each other; and two gDMRs on 18q21, 2-78 and 3-8, also were adjacent to each other (Table 1). In addition, 14 methylated CpG islands were located near and on the same chromosomal subband as other methylated CpG islands (Table 1). For example, SMRs 3-110, 2-1, and 1-12 are all on 17q25; two of these sequences, 3-110 and 1-12, lie within 660 kb. In some cases, SMRs and gDMRs were found in relatively close proximity. For example, SMR 2-3 and gDMR 1-13 lie within 1 Mb on 18q23. In addition, gDMR 1-20 and SMR 3-12 are both on 10q26 and separated by ~800 kb (Table 1). All of these data together support the idea that these methylated CpG islands identify specific portions of the genome.

[**0161**] Most of the methylated CpG islands were localized within or near the coding sequence of known genes or of anonymous ESTs within the GenBank or Celera databases. Because of the known ability of DMRs to regulate imprinting over long distances (reviewed in Feinberg 2001), the identity of known or predicted genes within several hundred kilobases of each methylated CpG rich region, was determined. Particularly intriguing was the discovery that gDMR 3-4 was located on 6q24 within HYMA1 (**FIG. 5**), an imprinted gene involved in diabetes mellitus (Arima et al. 2000). This CpG rich region has been identified independently as a DMR, in a specific analysis of this gene (Arima et al. 2001), and isolation of this sequence using a method of the present invention indicates that these methylated CpG islands may identify imprinted gene domains. gDMR 1-13 was located on 18q23, within a predicted gene of unknown function, and near the SALL3 gene (**FIG. 5**), which encodes a Spalt-like zinc finger protein that is a candidate gene for 18q deletion syndrome (10610715), which involves preferential loss of the paternal allele (Kohlhase et al. 1999). Interestingly, 18q23 also has been implicated in bipolar affective disorder, specifically harboring a predisposing gene transmitted preferentially through the father (Stine et al. 1995; McMahon et al. 1997). Therefore, the localization of this gDMR may serve as a guidepost for identifying candidate imprinted genes for this important disease. SMR 1-2 was located within 19q 13.4 (**FIG. 5**). Even though this sequence is an SMR, 19q13.4 contains the imprinted genes PEG3 and ZIM1 (Kim et al. 1999). Given that SMR 1-2 is ~10 Mb from these genes, it is unlikely to lie within the same imprinted gene domain. Nevertheless, it will be of interest to examine nearby genes for their imprinting status, including a glioma tumor suppressor candidate gene located 110 kb telomeric to SMR 1-2. Another interesting gene harboring a methylated CpG rich region was histone deacetylase A (HDAC4), and there were several other predicted genes near this CpG rich region, SMR 3-20 (**FIG. 5**). In addition, several antisense transcripts are associated with this CpG rich region. Given that HDAC4 is itself involved in chromatin remodeling (Wang et al. 2000), methylation of this region could be involved in a feedback loop controlling chromatin structure. Other genes located near methylated CpG islands included the wolframin gene, a transmembrane protein involved in congenital diabetes (Strom et al. 1998); several olfactory receptor genes; several phosphatase and kinase genes likely involved in signal transduction; several genes for DNA-interacting proteins; and the Peutz-Jeghers syndrome gene STK11 (Table 1). A voltage-dependent potassium channel subunit protein was localized only 16 kb

from methylated CpG rich region 2-3 (Table 1), which is of interest given that the voltage-dependent potassium channel KvLQT1 is imprinted (Lee et al. 1997). Finally, in addition to genes directly adjacent to these methylated CpG islands, at least two of the domains flanked by methylated CpG islands harbored several genes within them that may play a role in cancer. For example, contained within the region defined by methylated CpG islands 3-110 and 1-12 are a predicted apoptosis inhibitor, a septin-like cell division gene, a ras homolog, and a predicted translation initiation factor (Table 1).

EXAMPLE 4

[0162] Identification of an Imprinted Gene Homologous to Elongin A

[0163] This example illustrates the use of the methods of the present invention for identifying novel genes associated with CpG islands. More specifically, this example illustrates the use of the methods of the present invention to identify the Elongin A gene.

[0164] Imprinting Analysis of Elongin A3 Gene

[0165] Fetal tissues and matched maternal decidua were obtained from the University of Washington Fetal Tissue Bank. We identified polymorphisms by sequencing fetal and maternal PCR amplified genomic DNA. The following conditions were used for PCR amplifications: 95° C., 2 min; then 40 cycles of 95° C. 1 min, 60° C. 30 sec, 72° C. min; then 72° C. for 9 min. Total RNA was isolated from fetal tissues using RNeasy mini kit (Qiagen). To eliminate DNA contamination from RNA preparations, samples were treated with preamplification-grade DNase I (Invitrogen) according to supplied protocols. RT-PCR was carried out using the Superscript II preamplification system (Invitrogen) and was performed for each sample in the presence and absence (negative controls) of RT. cDNA samples were sequenced only when no bands were obtained with the negative controls. The primers used for the imprinting analysis were EL2AL-1093-1112F: 5'-TCT GCT GTC CGC TTT TGA GG-3' (SEQ ID NO: 32) and EL2AL-1526-1550R: 5'-ATC GGA TTT TCG TGG TCA CTA CTC G-3' (SEQ ID NO: 33). DNA and cDNA sequencing was run on an ABI-377 automated sequencer following protocols recommended by the manufacturer (Perkin-Elmer).

[0166] In addition to HYMA1, described above, a DMR within the IGF2R contains an Eag I site, and as predicted, this gene also was found in the Eag library. Allele-specific expression of genes near methylated islands was examined. gDMR 2-78 was localized to 18q21 (**FIG. 5**) and was completely methylated in all somatic fetal and adult tissues tested (**FIG. 2**). However, this CpG rich region was unmethylated in CHM and sperm and methylated in OT (**FIG. 3A**). A BLAST search showed that the CpG rich region spanned the putative promoter region and body of a gene predicted by GENSCAN (http://genes.mit.edu/GENSCAN), and included 1638 nucleotides encoding 546 amino acids (**FIG. 6**). BLAST searches of GenBank and Celera databases using the predicted sequences revealed that the predicted gene showed 43% amino acid identity to human transcription elongation factor B (SIII) polypeptide 3 (TCEB3), also known as Elongin A. The novel sequence was even more closely related to a previously identified homolog of Elongin A termed Elongin A2, or TCEB3L, showing 79% amino acid sequence identity to human transcription elongation factor (SIII) Elongin A2 (TCEB3L). To determine whether 2-78 represented a genuine transcript, and if so, whether the gene is imprinted, primers were designed that would amplify 2-78 but not Elongin A2, and amplification products were of the expected size. Sequencing demonstrated that the amplified cDNA corresponded to 2-78 and not Elongin A2, based on sequence differences between the two genes within the PCR product. Analyzing DNA samples from fetal tissues, we then identified a polymorphism at nucleotide 910 (G/A) of 2-78. Four fetuses heterozygous for this polymorphism were identified, in which maternal decidua DNA was available and homozygous, allowing the identification of parental origin in the fetal samples (**FIG. 7**). Reverse transcriptase PCR (RT-PCR) analysis of tissues from these fetuses showed that the gene was indeed transcribed. We therefore term this gene Elongin A3. An alternative term is TCEB3L2, but for this term to apply, the nomenclature committee will need to rename TCEB3L (Elongin A2) TCEB3L1.

[0167] Analysis of allele-specific expression showed monoallelic expression of lung, brain, placenta, and spinal cord, with preferential expression from the maternal allele (FIGS. 7A-D). There was incomplete preferential expression from the maternal allele in two of three kidneys (**FIGS. 7A, C**), and absence of imprint-specific gene expression in one kidney and in the intestine or liver (**FIGS. 7B, C**, D). Thus, Elongin A3 shows tissue-specific imprinting, at least in prenatal development. Therefore, the isolation of these novel CpG islands does enable the identification of novel human imprinted genes.

EXAMPLE 5

[0168] Species Conservation of Methylated CpG Rich Regions

[0169] This example illustrates that the CpG islands identified herein are conserved among mammalian species and can be used to identify nearby regulatory elements conserved between species.

[0170] As further confirmation of the importance of the methylated CpG islands that were isolated, their sequence conservation in the mouse was ascertained using the Celera mouse genome database. Thirteen (46%) of the 30 human noncontiguous methylated CpG islands matched sequences within the mouse genome at 86.9±4.9% identity (**FIG. 8**). Furthermore, in some cases, the region of conservation extended beyond the CpG rich region itself. For example, gDMR 1-21 showed, in addition to a 558 bp, 82% conserved region including the CpG rich region, five additional conserved sequences within 1 kb of the CpG rich region. These additional sequences varied from 80-97% identity (**FIG. 8**). Most of the conserved sequences outside of the CpG islands themselves were not predicted genes, and thus may represent conserved regulatory sequences. In all cases in which BLAST analysis of the CpG rich region and flanking 1 kb on each side was performed, and in which any sequence conservation was found, the CpG rich region itself was conserved, again supporting the idea that these CpG islands play an important role.

EXAMPLE 6

[0171] Normally Methylated CpG Islands and GC Rich Sequences

[0172] This Example provides further insight into the methods of the present invention and the conclusions

reported herein. A major conclusion of the previous Examples is the identification of a subset of unique CpG islands that are methylated in normal tissues, in the first systematic effort to identify such sequences. The experiments were designed to identify CpG islands that are methylated differentially in germline-derived tissues or differentially in cancers. However, no CpG islands methylated specifically in tumors were found, but slightly more than one half of the unique methylated CpG islands were methylated in germline-derived tissues of both maternal and paternal origin. Conventional wisdom holds that CpG islands are unmethylated, with the exception of the inactive X chromosome, imprinted genes, and tumors. However, rare exceptions to this rule have been described. Some repeated sequences harboring CpG islands have been found to be methylated. Methylation of a mouse testis-specific histone H2B gene has been reported (Choi et al. 1996), and others have found methylation of some ribosomal gene sequences (Brock and Bird 1997). Indeed, methylation of one of these repeat sequences, the rDNA nontranscribed spacer, previously was found after genomic purification from a methyl-CpG binding protein column (Brock and Bird 1997), and the large number of these sequences may have obscured the identification of unique methylated CpG islands. The methylation of high copy number sequences is not surprising, as it is consistent with the hypothesis that CpG methylation arose as a host defense mechanism (Bestor and Tycko 1996). This is particularly true of the SVA element, which is a high copy number retroposon.

[0173] However, the presence of normally methylated unique CpG islands and GC rich regions has not been observed systematically. An intriguing exception is the MAGE melanoma gene (Serrano et al. 1996), and it is thought that hypomethylation of this gene leads to its activation in cancer (De Smet et al. 1996). Our results suggest that normally methylated single-copy CpG islands and GC rich regions may be more abundant than previously believed. Indeed, the loss of methylation of such sequences may be related to gene activation in cancer, just as the gain of methylation of CpG islands and GC rich regions may lead to their silencing. Previous screens for altered CpG rich region methylation have not been designed to identify normally methylated CpG islands and GC rich regions, but it should be noted that the original observation of altered methylation in cancer was widespread loss of methylation (Feinberg and Vogelstein 1983). Furthermore, even in tumors that show increased CpG rich region methylation, the total methylation content is reduced (Feinberg et al. 1988). DNA methylation serves as an additional layer of genetic information in the genome, which has been termed the methylome (Feinberg 2001), and both increases and decreases may be important in cancer. Our strategy for cloning these sequences can be generalized to secondary libraries in addition to the Eag library, and the identification of additional such sequences thus should enhance our understanding of the methylome.

[0174] Another major result of the above Examples is the identification of novel CpG islands and GC rich regions that are methylated differentially in OT and CHM. The second (Eag) library would not identify known imprinted genes lacking Eag I sites, but it did contain the DMR of IGF2R (Wutz and Barlow, Mol. Cell Endocrinol. May 25, 1988;140(1-2):9-14), as well as the DMR of the imprinted HYMA1 gene, suggesting that this strategy also can identify novel imprinted gene domains. One such gene was identified to date, a novel homolog of the Elongin A and Elongin A2 genes, which we term Elongin A3. Both Elongin A and Elongin A2 are known to be the active components of the transcription factor B (SIII) complex (Aso et al. 2000), that may compete for other components (Elongin B and C) with the VHL tumor suppressor gene (Kibel et al. 1995). We did not check directly for elongation activity of Elongin A3, but it contains the TFS2N motif as well as a nuclear localization signal, and the predicted protein sequence is 79% identical to that of Elongin A2, so it likely does have such a function.

[0175] It should be noted that gDMRs, even the gDMR within this novel imprinted gene, showed variable to complete methylation in somatic tissues. Such a pattern of methylation also is similar to that seen for the promoter of the imprinted gene ZNF127 (Strom et al. 1998), and for at least one methylated CpG rich region within the 11p15 imprinted gene domain. Thus, imprinted gene domains may harbor some methylated CpG islands and GC rich regions that show persistent differential methylation in somatic tissues, but also may contain other CpG islands and GC rich regions that do not show these differences in somatic tissues. Thus, it is important to compare methylation in sperm or CHM as a representation of the male germline, and OT (as eggs cannot be harvested from humans for this purpose), in the search for imprinted gene domains. The mouse is a useful adjunct and provides access to a greater variety of tissues at varying developmental stages, but there are substantial differences between human and mouse imprinting, both in the identity of the genes themselves, and in their developmental pattern of imprinting.

[0176] Several of these domains harbor multiple genes that have been implicated in cancer, and that show frequent loss of heterozygosity, including 4p16, 4q35, 10q26, 18q21, and 19p13. An imprinted tumor suppressor gene in one or more of these regions might not show conventional mutations in tumors, and thus identifying imprinted genes is an important part of tumor suppressor gene identification within these regions. The same region of 18q also has shown linkage in bipolar affective disorder, with preferential transmission through the paternal allele (McMahon et al. 1997). Furthermore, these domains appear to harbor both SMRs and gDMRs, suggesting that both types of methylated CpG islands and GC rich regions may be useful for identifying imprinted gene domains.

[0177] CpG islands and GC rich regions normally must be under selective pressure for their maintenance, as methylation leads to deamination and loss of cytosine. This is especially true in the case of the SMRs we have described, as they are methylated even in sperm DNA. In the case of gDMRs, their methylation in somatic tissues and oocyte-derived cells may be critical for suppression of nearby gene expression in spermatocyte progenitor cells. This may be particularly important for genes involved in establishing epigenetic states and in epigenetic reprogramming, as the chromatin of spermatocyte differs markedly from oocytes and somatic cells.

[0178] It also is likely that normally methylated CpG islands and GC rich regions are involved directly in chromatin formation. For example, they could serve as chromatin insulators separating enhancers from promoters. If that is so, then we would expect to find their loss of methylation in

specific tissues at specific developmental stages, which would be consistent with the observation that imprinted genes can show developmental (tissue- and timing-specific) imprinting (Lee et al. 1997). Support for this idea also comes from our observation that SMRs were more frequently localized near the ends of chromosomes. Given that chromosomal ends are associated with the nuclear lamina in interphase (Cockell and Gasser 1999), the relative proximity of SMRs to the ends of chromosomes might permit their association with the nuclear lamina and chromatin proteins found within it.

[0179] Normally methylated CpG islands and GC rich regions also might promote chromatin formation. In an intriguing review, Pardo-Manuel de Villena et al. (2000) suggest that imprinting involving differences among homologous chromosomes arose under selective pressure to facilitate pairing and distinguish homologous chromosomes during meiosis. We suggest that SMRs also might enhance pairing and recombination by recruiting chromatin factors to specific locations along a given chromosome and allowing those factors to interact between homologous chromosomes. A prediction of our suggestion is that recombination frequencies in meiosis or even mitosis might be enhanced near normally methylated CpG islands and GC rich regions. Methylated CpG islands and GC rich regions also may play a role intrachromosomal compartmentalization. For example, the gDMRs lay within regions of comparatively lower CpG content (GC-poor isochores). Consistent with this idea, we have noted that most known imprinted genes also appear to lie within low isochore regions (PLAGL1, IGF2R, PEG1/MEST, SNRPN, PEG3, GNAS).

[0180] Finally, the identification of these methylated CpG islands and GC rich regions will facilitate comparison of their sequences to each other, as well as computational analysis of sequence motifs. For example, in preliminary experiments, several CTCF binding sites within at least 10 methylated CpG islands and GC rich regions have been identified. Therefore, CTCF binding may be a common feature of these sequences.

[0181] It has recently been proposed that CpG islands and GC rich regions fall into several groups, one of which represents unique CpG and generally unmethylated islands and GC rich regions associated with the 5' region of housekeeping genes, whereas another includes high-copy nongene CpG islands and GC rich regions that are dominated by Alu I repeat elements (Ponger et al. 2001). Because Alu I repeats are generally methylated and transcriptionally silent, high-copy CpG islands and GC rich regions are predicted to be methylated. Indeed, the report of Strichman-Almashanu et al. (2002) identified one of the high-copy CpG islands and GC rich regions (SVA) to be heavily methylated. This observation is not surprising given that repeat sequences provide signatures for de novo methylation, according to the host defense model (Bestor and Tycko 1996).

[0182] Strichman-Almashanu et al. (2002) also report the existence of a new class of unique CpG islands and GC rich regions that are methylated on both alleles in all tissues examined. Interestingly, these CpG islands and GC rich regions (SMRs) mapped to isochores with high GC content (>0.5), whereas the differentially methylated islands and GC rich regions (gDMRs) were concentrated in isochores with low GC content (<0.5). The class of unmethylated or differentially methylated CpG islands and GC rich regions could stand out in a CpG-less environment and provide landmarks for various recognition events, such as the initiation of chromatin condensation by TP2 during spermiogenesis (Kundu and Rao 1996). The complexity of CpG rich region compartmentalization of the mammalian genome was further emphasized by the observation that the methylated high-copy CpG islands and GC rich regions frequently localize close to telomeric ends (Strichman-Almanshanu et al. 2002), as do densely methylated nonrich region CpG stretches (Brock et al. 1999), indicating some methylation-dependent role in chromosomal integrity. This deduction is supported by the observation that DNA methyltransferase, Dnmt1, is essential for genomic stability in mouse embryonic stem cells (Chen et al. 1998).

TABLE 3

Chromosomal Location of SMRs and gDMRs

Clone ID		Methylation pattern	Chromosomal location Accession no.
3-10	SMR	1q44 TEL	14042537a
			14423768
			hCG1644736
			hCG1724357
3-20	SMR	2q36 SUBTEL	5174481a
			hCG1651464
			hCG1656118
			hCG1651461
			hCG1651466
1-19	SMR	4p16 TEL	3777583a
1-41	SMR	4q35 TEL	hCG1788598a
			hCG1793025a
			hCG1787540
4-8	SMR	4q35 TEL	hCG1788598a
3-4	gDMR	6q24 SUBTEL	hCG1660630a
4-7	SMR	7p22 TEL	6806913a
			hCG1747708
			hCG1790856
			hCG1747710
1-30	SMR	7q11.1	hCG1779529a
			hCG1779527
			hCG1789113
			13642872
			6572672
1-22	SMR	7q36 TEL	11386149a
			hCG1799787
3-2	SMR	8p23 TEL	hCG1659058a
2-5	gDMR	8q21.2	17451956
			hCG1757665
2-48	gDMR	9p13	hCG1659616
1-20	gDMR	10q26 TEL	13325182a
			hCG1799063
3-12	SMR	10q26 TEL	3122245a
			hCG1654478
3-30	gDMR	11q25 TEL	17456499a
			hCG37607
			hCG1745526
1-5	SMR	13q34 TEL	hCG20146a
1-21	gDMR	14q32 TEL	8393715a
			hCG21408
2-42	gDMR	16p13.1	hCG15669a
3-110	SMR	17q25 TEL	8400736a
			10435982
2-1	SMR	17q25 TEL	1655842a
			14149793
1-12	SMR	17q25 TEL	hCG1806389
			hCG1796817
2-78	gDMR	18q21	13645769a
3-8	gDMR	18q21	13645769a
2-3	SMR	18q23 TEL	6912444
			1914872
			5326898

TABLE 3-continued

Chromosomal Location of SMRs and gDMRs

	Clone ID	Methylation pattern	Chromosomal location Accession no.
1-13	gDMR	18q23 TEL	hCG1651089a
			6688241
			hCG20372
			1651088
1-6	gDMR	19p13.1	14249150a
4-3	SMR	19p13.3 TEL	9665054a
			hCG23965
			4506715
			hCG1794585
			10732648
1-2	SMR	19q13.4 TEL	12053197
			4505329
			5901994
			5689511
			7657128
			7657054
			7657130
2-4	gDMR	20q12	17484155a
			hCG1800975
			hCG1653833
			13378306
			110743
			7799072

[0183] References

[0184] Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. 1990. Basic local alignment search tool. J. Mol. Biol. 215: 403-410.

[0185] Antequera, F. and Bird, A. P. 1993. Number of CpG rich regions and genes in human and mouse. Proc. Natl. Acad. Sci. 90: 11995-11999.

[0186] Arima, T., Drewell, R. A., Arney, K. L., Inoue, J., Makita, Y., Hata, A., Oshimura, W., Wake, N., and Surani, M. A. 2001. A conserved imprinting control region at the HYMAI/ZAC domain is implicated in transient neonatal diabetes mellitus. Hum. Mol. Genet. 10:1475-1483.

[0187] Arima, T., Drewell, R. A., Oshimura, M., Wake, N., and Surani, M. A. 2000. A novel imprinted gene, HYMAI, is located within an imprinted domain on human chromosome 6 containing ZAC. Genomics 67: 248-255.

[0188] Aso, T., Yamazaki, K., Amimoto, K., Kuroiwa, A., Higashi, H., Matsuda, Y., kitajuma, S., and Hatakeyama, M. 2000. Identification and characterization of Elongin A2, a new member of the Elongin family of transcription elongation factors, specifically expressed in the testis. J. Biol. Chem. 275: 6546-6552.

[0189] Barlow, D. P. 1995. Gametic imprinting in mammals. Science 270: 1610-1613.

[0190] Bernardi, G. 1995. The human genome: Organization and evolutionary history. Ann. Rev. Genet. 29: 445-476.

[0191] Bestor, T. H. and Tycko, B. 1996. Creation of genomic methylation patterns. Nat. Genet. 12: 363-367.

[0192] Bird, A. P. 1986. CpG-rich rich regions and the function of DNA methylation. Nature 321: 209-213.

[0193] Bird, A. P., Taggart, M., Frommer, M., Miller, O. J., and Macleod, D. 1985. A fraction of the mouse genome that is derived from rich regions of nonmethylated, CpG-rich DNA. Cell 40: 91-99.

[0194] Brock, G. J., Charlton, J., and Bird, A. P. 1999. Densely methylated sequences that are preferentially localized at telomere-proximal regions of human chromosomes. Gene 240: 269-277.

[0195] Brock, G. J. R. and Bird, A. P. 1997. Mosaic methylation of the repeat unit of the human ribosomal RNA genes. Hum. Mol. Genet. 6: 451-456.

[0196] Choi, Y. -C., Gu, W., Hecht, N. B., Feinberg, A. P., and Chae, C. -B. 1996. Molecular cloning of mouse somatic and testis-specific H2B histone genes containing a methylated CpG rich region. DNA Cell Biol. 15: 495-504.

[0197] Cockell, M. and Gasser, S. M. 1999. Nuclear compartments and gene regulation. Curr. Opin. Genet. Dev. 9: 199-205.

[0198] Cross, S. H. and Bird, A. P. 1995. CpG rich regions and genes. Curr. Opin. Genet. Dev. 5: 309-314.

[0199] De Smet, C., De Backer, O., Faraoni, I., Lurquin, C., Brasseur, F., and Boon, T. 1996. The activation of human gene MAGE-1 in tumor cells is correlated with genome-wide demethylation. Proc. Natl. Acad. Sci. 93: 7149-7153.

[0200] Dyson, N. J. 1991. Essential molecular biology: A practical approach (ed. T. A. Brown), Vol. 2, pp. 111-156. IRL Press, Oxford.

[0201] Feinberg, A. P. 2001. Cancer epigenetics takes center stage. Proc. Natl. Acad. Sci. 98: 392-394.

[0202] Feinberg, A. P. 2001. Methylation meets genomics. Nat. Genet. 27: 9-10.

[0203] Feinberg, A. P. and Vogelstein, B. 1983. Hypomethylation distinguishes genes of some human cancers from their normal counterparts. Nature 301: 89-92.

[0204] Feinberg, A. P., Gehrke, C. W., Kuo, K. C., and Ehrlich, M. 1988. Reduced genomic 5-methylcytosine content in human colonic neoplasia. Cancer Res. 48: 1159-1161.

[0205] Ferguson-Smith, A. C., Sasaki, H., Cattanach, B. M., and Surani, M. A. 1993. Parental-origin-specific epigenetic modification of the mouse H19 gene. Nature 362: 751-755.

[0206] Gardiner-Garden, M. and Frommer, M. 1987. CpG rich regions in vertebrate genomes. J. Mol. Biol. 196: 261-282.

[0207] Hayashizaki, Y., Shibata, H., Hirotsune, S., Sugino, H., Okazaki, Y., Sasaki, N., Hirose, K., Imoto, H., Okuizumi, H., Muramatsu, M. 1994. Identification of an imprinted U2af binding protein related sequence on mouse chromosome 11 using the RLGS method. Nat. Genet. 6: 33-40.

[0208] Herman, J. G., Latif, F., Weng, Y., Lerman, M. I., Zbar, B., Liu, S., Samid, D., Duan, D. R., Gnarra, G. R.,

Linehan, W. M. 1994. Silencing of the VHL tumor-suppressor gene by DNA methylation in renal carcinoma. Proc. Natl. Acad. Sci. 91: 9700-9704.

[0209] Huang, T. H., Perry, M. R., and Laux, D. E. 1999. Methylation profiling of CpG rich regions in human breast cancer cells. Hum. Mol. Genet. 8: 459-470.

[0210] Kawajiri, K., Watanabe, J., Gotoh, O., Tagashira, Y., Sogawa, K., and Fujii-Kuriyama, Y. 1986. Structure and drug inducibility of the human cytochrome P-450c gene. Eur. J. Biochem. 159: 219-225.

[0211] Kibel, A., Iliopoulos, O., DeCaprio, J. A., and Kaelin, W. G., Jr. 1995. Binding of the von Hippel-Lindau tumor suppressor protein to Elongin B and C. Science 269: 1444-1446.

[0212] Kim, J., Lu, X., and Stubbs, L. 1999. Zim1, a maternally expressed mouse Kruppel-type zinc-finger gene located in proximal chromosome 7. Hum. Mol. Genet. 8: 847-854.

[0213] Kohlhase, J., Hausmann, S., Stojmenovic, G., Dixkens, C., Bink, K., Schulz-Schaeffer, W., Altmann, M., and Engel, W. 1999. SALL3, a new member of the human spalt-like gene family, maps to 18q23. Genomics 62: 216-222.

[0214] Larsen, F., Gundersen, G., Lopez, R., and Prydz, H. 1992. CpG rich regions as gene markers in the human genome. Genomics 13: 1095-1107.

[0215] Lee, M. P., Hu, R. -J., Johnson, L. A., and Feinberg, A. P. 1997. Human KVLQT1 gene shows tissue-specific imprinting and encompasses Beckwith-Wiedemann syndrome chromosomal rearrangements. Nat. Genet. 15: 181-185.

[0216] McMahon, F. J., Hopkins, P. J., Xu, J., McInnis, M. G., Shaw, S., Cardon, L., Simpson, S. G., MacKinnon, D. F., Stine, O. C., Sherrington, R. 1997. Linkage of bipolar affective disorder to chromosome 18 markers in a new pedigree series. Am. J. Hum. Genet. 61: 1397-1404.

[0217] Merlo, A., Herman, J. G., Mao, L., Lee, D., Gabrielson, E., Burger, P. C., Baylin, S. B., and Sidransky, D. 1995. 5' CpG rich region methylation is associated with transcriptional silencing of the tumour suppressor p16/CDKN2/MTS1 in human cancers. Nat. Med. 1: 686-692.

[0218] Ohlsson, R., Renkawitz, R., and Lobanenkov, V. 2001. CTCF is a uniquely versatile transcription regulator linked to epigenetics and disease. Trends Genet. 17: 520-527.

[0219] Pardo-Manuel de Villena, F., de la Casa-Esperon, E., and Sapienza, C. 2000. Natural selection and the function of genome imprinting: Beyond the silenced minority. Trends Genet. 16: 573-579.

[0220] Plass, C., Shibata, H., Kalcheva, I., Mullins, L., Kotelevtseva, N., Mullins, J., Kato, R., Sasaki, H., Hirotsune, S., Okazaki, Y. 1996. Identification of Grfl on mouse chromosome 9 as an imprinted gene by RLGS-M. Nat. Genet. 14: 106-109.

[0221] Razin, A. and Cedar, H. 1994. DNA methylation and genomic imprinting. Cell 77: 473-476.

[0222] Serrano, A., Garcia, A., Abril, E., Garrido, F., and Ruiz-Cabello, F. 1996. Methylated CpG points identified within MAGE-1 promoter are involved in gene repression. Int. J. Cancer 68: 464-470.

[0223] Shen, L., Wu, L. C., Sanlioglu, S., Chen, R., Mendoza, A. R., Dangel, A. W., Carroll, M. C., Zipf, W. B., and Yu, C. Y. 1994. Structure and genetics of the partially duplicated gene RP located immediately upstream of the complement C4A and the C4B genes in the HLA class III region. Molecular cloning, exon-intron structure, composite retroposon, and breakpoint of gene duplication. J. Biol. Chem. 269: 8466-8476.

[0224] Shiraishi, M., Chuu, Y. H., and Sekiya, T. 1999. Isolation of DNA fragments associated with methylated CpG rich regions in human adenocarcinomas of the lung using a methylated DNA binding column and denaturing gradient gel electrophoresis. Proc. Natl. Acad. Sci. 96: 2913-2918.

[0225] Stine, O. C., Xu, J., Koskela, R., McMahon, F. J., Gschwend, M., Friddle, C., Clark, C. D., McInnis, M. G., Simpson, S. G., and Breschel, T. S. 1995. Evidence for linkage of bipolar disorder to chromosome 18 with a parent-of-origin effect. Am. J. Hum. Genet. 57: 1384-1394.

[0226] Strom, T. M., Hortnagel, K., Hofmann, S., Gekeler, F., Scharfe, C., Rabl, W., Gerbitz, K. D., and Meitinger, T. 1998. Diabetes insipidus, diabetes mellitus, optic atrophy and deafness (DIDMOAD) caused by mutations in a novel gene (wolframin) coding for a predicted transmembrane protein. Hum. Mol. Genet. 7: 2021-2028.

[0227] Toyota, M., Ho, C., Ahuja, N., Jair, K. -W., Li, Q., Ohe-Toyota, M., Baylin, S. B., and Issa, J. -P. J. 1999. Identification of differentially methylated sequences in colorectal cancer by methylated CpG rich region amplification. Cancer Res. 59: 2307-2312.

[0228] Wang, A. H., Kruhlak, M. J., Wu, J., Bertos, N. R., Vezmar, M., Posner, B. I., Bazett-Jones, D. P., and Yang, X. J. 2000. Regulation of histone deacetylase 4 by binding of 14-3-3 proteins. Mol. Cell. Biol. 20: 6904-6912.

[0229] Yen, P. H., Patel, P., Chinault, A. C., Mohandas, T., and Shapiro, L. 1984. Differential methylation of hypoxanthine phosphoribosyltransferase genes on active and inactive human X chromosomes. Proc. Natl. Acad. Sci. 81: 1759-1763.

[0230] Zhu, Z. B., Hsieh, S., Bently, D. R., Campbell, D. R., and Volanakis, J. E. 1992. A variable number of tandem repeats locus within the human complement C2 gene is associated with a retroposon derived from a human endogenous retrovirus. J. Exp. Med. 175: 1783-1787.

[0231] Bestor, T. and Tycko, B. 1996. Nat. Genet. 12: 363-367[Medline].

[0232] Brandeis, M., Frank, D., Keshet, I., Siegfried, Z., Mendelsohn, M., Nemes, A., Temper, V., Razin, A., and Cedar, H. 1994. Nature 371: 435-438[Medline].

[0233] Brock, G., Charlton, J., and Bird, A. 1999. Gene 240: 269-277[CrossRef][Medline].

[0234] Chen, R., Pettersson, U., Beard, C., Jackson-Grusby, L., and Jaenisch, R. 1998. Nature 395: 89-93 [CrossRef][Medline].

[0235] Cross, S. and Bird, A. 1995. Curr. Opin. Genet. Dev. 5: 309-314[Medline].

[0236] De Smet, C., De Backer, O., Faraoni, I., Lurquin, C., Brasseur, F., and Boon, T. 1996. Proc. Natl. Acad. Sci. 93: 7149-7153[Abstract].

[0237] Feinberg, A. and Vogelstein, B. 1983. Nature 301: 89-92[Medline].

[0238] Hejnar, J., Hajkova, P., Plachy, J., Elleder, D., Stepanets, V., and Svoboda, J. 2001. Proc. Natl. Acad. Sci. USA 98: 565-569[Abstract/Full Text].

[0239] Issa, J. -P. J. and Baylin, S. B. 1996. Nat. Med. 2: 281-282[Medline].

[0240] Kundu, T. and Rao, M. 1996. Biochemistry 35: 15626-15632[CrossRef][Medline].

[0241] Macleod, D., Charlton, J., Mullins, J., and Bird, A. P. 1994. Genes Dev. 8: 2282-2292[Abstract].

[0242] Ohlsson, R., Cui, H., He, L., Pfeifer, S., Jiang, S., Feinberg, A. P., and Hedborg, F. 1999. Cancer Res. 59: 3889-3892[Abstract/Full Text].

[0243] Ohlsson, R., Renkawitz, R., and Lobanenkov, V. 2001. Trends Genet. 17: 520-527[Medline].

[0244] Ponger, L., Duret, L., and Mouchiroud, D. 2001. Genome Res. 11: 1854-1860[Abstract/Full Text].

[0245] Razin, A. and Cedar, H. 1994. Cell 77: 473-476 [Medline].

[0246] Strichman-Almanshanu, L., Lee, R., Onyango, P., Perlman, E., Flam, F., Frieman, M., and Feinberg, A. 2002. Genome Res. X, Y. (incorporated herein by reference in its entirety).

[0247] Voo, K. S, Carlone, D. L., Jacobsen, B. M., Flodin, A., and Skalnik, D. G. 2000. Mol. Cell. Biol. 20: 2108-2121[Abstract/Full Text].

[0248] Yan, P., Chen, C., Shi, H., Rahmatpanah, F., Wei, S., Caldwell, C., and Huang, T. 2001. Cancer Res. 61: 8375-8380[Abstract/Full Text].

[0249] Although the invention has been described with reference to the above examples, it will be understood that modifications and variations are encompassed within the spirit and scope of the invention. Accordingly, the invention is limited only by the following claims.

What is claimed is:

1. A method for determining a disease state in a subject comprising:

determining the DNA methylation status at a cytosine residue of a CpG sequence in a genomic DNA sample from the subject, wherein hypomethylation of a CpG sequence normally methylated in a subject not having the disease state, is indicative of a disease state in the subject.

2. The method of claim 1, wherein the CpG sequence is within a GC rich region or a CpG island.

3. The method of claim 1, wherein the subject is a human.

4. The method of claim 1, wherein the disease is cancer.

5. The method of claim 1, wherein the disease is selected from cancer, multiple sclerosis, Alzheimer's disease, Parkinson's disease, depression and other imbalances of mental stability, atherosclerosis, cystic fibrosis, diabetes, obesity, Crohn's disease, and altered circadian rhythmicity, arthritis, inflammatory reactions or disorders, psoriasis and other skin diseases, autoimmune diseases, allergies, hypertension, anxiety disorders, schizophrenia and other psychoses, osteoporosis, muscular dystrophy, amyotrophic lateral sclerosis or circadian rhythm-related conditions.

6. A method for determining the DNA methylation status at a cytosine residue of a CpG sequence in a genomic DNA sample, comprising performing methylation state analysis of one or more CpG islands or GC rich regions of a genomic DNA sample, thereby determining the DNA methylation status in the genomic DNA sample.

7. The method of claim 6, wherein the one or more CpG islands or GC rich regions include differentially methylated regions (DMRs).

8. The method of claim 6, wherein the one or more CpG islands and GC rich regions include similarly methylated regions (SMRs).

9. The method of claim 6, wherein the CpG island or GC rich region is selected from the group consisting of SEQ ID NOs: 3-7, SEQ ID NO: 9-30 and SEQ ID NO: 31 and any combination thereof.

10. The method of claim 8, wherein the CpG island or GC rich region includes at least one of SEQ ID NO: 3, SEQ ID NO: 4, SEQ ID NO: 5, SEQ ID NO: 6, SEQ ID NO: 7, SEQ ID NO: 9, SEQ ID NO: 10, SEQ ID NO: 11, SEQ ID NO: 12, SEQ ID NO: 16, SEQ ID NO: 21, SEQ ID NO: 22, SEQ ID NO: 23, SEQ ID NO: 26, SEQ ID NO: 29, or SEQ ID NO: 30.

11. The method of claim 8, wherein the CpG island or GC rich region includes at least SEQ ID NO: 4.

12. The method of claim 7, wherein the CpG islands and GC rich regions are least one of SEQ ID NO: 13, SEQ ID NO: 14, SEQ ID NO: 15, SEQ ID NO: 17, SEQ ID NO: 19, SEQ ID NO: 20, SEQ ID NO: 24, SEQ ID NO: 25, SEQ ID NO: 27, SEQ ID NO: 28, and SEQ ID NO: 31.

13. The method of claim 6, wherein the methylation state of at least two CpG islands and GC rich regions is identified.

14. The method of claim 6, wherein the methylation state of at least three CpG islands and GC rich regions is identified.

15. The method of claim 6, wherein the methylation state of SEQ ID NO: 27 is identified.

16. The method of claim 6, wherein the methylation state of SEQ ID NO: 30 is identified.

17. The method of claim 6, wherein the methylation state of SEQ ID NO: 24 is identified.

18. The method of claim 6, wherein the methylation state of SEQ ID NO: 8 is identified.

19. The method of claim 6, wherein the methylation state of SEQ ID NO: 4 is identified.

20. The method of claim 6, wherein the methylation state of SEQ ID NO: 26 is identified

21. The method of claim 6, wherein the methylation state of SEQ ID NO: 21 is identified.

22. The method of claim 6, wherein the methylation state of SEQ ID NO: 23 is identified.

23. The method of claim 6, wherein the methylation state of SEQ ID NO: 21 and SEQ ID NO: 23 is identified.

24. An isolated polynucleotide comprising about 1638 nucleotides encoding about 546 amino acids and having about 79% amino acid sequence identity to Elongin A2 and a Genbank accession number NM_145653.

25. The polynucleotide of claim 24, wherein nucleotide 910 is G or A.

26. The polynucleotide of claim 24, wherein the polynucleotide is set forth in SEQ ID NO: 1.

27. The polynucleotide of claim 24, wherein the polynucleotide encodes a polypeptide as set forth in SEQ ID NO: 2.

28. An isolated polynucleotide comprising SEQ ID NO: 1 and sequences 5' and 3' to SEQ ID NO: 1 containing CpG islands and GC rich regions.

29. A purified polypeptide encoded by the polynucleotide of claim 24.

30. Antibodies that bind to the polypeptide of claim 29.

31. A method for identifying the presence of an imprinted gene comprising:

comparing the DNA methylation status at a cytosine residue of a CpG sequence in a genomic DNA sample of maternal origin with the DNA methylation status at a cytosine residue of a CpG sequence in a genomic DNA sample of paternal origin, wherein a difference in DNA methylation status between the two samples is indicative of the presence of an imprinted gene.

32. The method of claim 31, wherein the difference in methylation status is in a GC rich region or a CpG island.

33. A methylation status representation identified by the method of claim 31.

34. A method for identifying the presence of an imprinted gene in genomic DNA comprising:

identifying a population of CpG islands and GC rich regions in a genomic DNA sample;

identifying a candidate gene within about 200 to 2000 kilobases of a first CpG island or GC rich region in the population of CpG islands and GC rich regions; and

determining whether the candidate gene is regulated by methylation of the first CpG island or GC rich region and preferentially methylated in paternal DNA or maternal DNA, wherein regulation of the candidate gene by methylation of the first CpG island or GC rich region and paternal or maternal preferential methylation is indicative of an imprinted gene, thereby identifying the presence of an imprinted gene.

35. An imprinted gene identified by the method of claim 34.

36. A method for identifying a CpG island or GC rich region-regulated gene, comprising identifying a candidate gene within about 200 to 2000 kilobases of a CpG island or GC rich region and determining whether the candidate gene is regulated by methylation of the CpG island or GC rich region, thereby identifying the CpG island or GC rich region-regulated gene.

37. The method of claim 34, wherein the method comprises determining whether the candidate gene is regulated by methylation of a CpG island or GC rich region is selected from SEQ ID NO: 3-31, with the proviso that it is not SEQ ID NO: 8.

38. The method of claim 34, wherein the candidate gene is regulated by methylation of SEQ ID NO: 13, SEQ ID NO: 14, SEQ ID NO: 15, SEQ ID NO: 17, SEQ ID NO: 19, SEQ ID NO: 20, SEQ ID NO: 24, SEQ ID NO: 25, SEQ ID NO: 27, SEQ ID NO: 28, or SEQ ID NO: 31.

39. The method of claim 34, wherein the method comprises determining whether the candidate gene is regulated by methylation of SEQ ID NO: 27.

40. The method of claim 34, wherein the method comprises determining whether the candidate gene is regulated by methylation of SEQ ID NO: 30.

41. The method of claim 34, wherein the method comprises determining whether the candidate gene is regulated by methylation of SEQ ID NO: 24.

42. The method of claim 34, wherein the method comprises determining whether the candidate gene is regulated by methylation of SEQ ID NO: 8.

43. The method of claim 34, wherein the method comprises determining whether the candidate gene is regulated by methylation of SEQ ID NO: 4.

44. The method of claim 34, wherein the method comprises determining whether the candidate gene is regulated by methylation of SEQ ID NO: 26.

45. The method of claim 34 wherein the method comprises determining whether the candidate gene is regulated by methylation of SEQ ID NO: 21.

46. The method of claim 34, wherein the method comprises determining whether the candidate gene is regulated by methylation of SEQ ID NO: 23.

47. The method of claim 34, wherein the method comprises determining whether the candidate gene is regulated by methylation of SEQ ID NO: 21 and SEQ ID NO: 23.

48. A method for determining the methylation status of a population of similarly methylated regions (SMRs) in a subject, comprising performing methylation status analysis of a population of SMRs of genomic DNA from a human sample.

49. The method of claim 48, wherein the methylation status of SMRs is correlated with a disease state.

50. The method of claim 48, wherein the population of SMRs is selected from at least two of SEQ ID NO: 3, SEQ ID NO: 4, SEQ ID NO: 5, SEQ ID NO: 6, SEQ ID NO: 7, SEQ ID NO: 9, SEQ ID NO:10, SEQ ID NO: 11, SEQ ID NO: 12, SEQ ID NO: 16, SEQ ID NO: 21, SEQ ID NO: 22, SEQ ID NO: 23, SEQ ID NO: 26, SEQ ID NO: 29, and SEQ ID NO: 30.

51. The method of claim 48, wherein the population of SMRs comprises three SMRs.

52. The method of claim 49, wherein the disease is selected from cancer, multiple sclerosis, Alzheimer's disease, Parkinson's disease, depression and other imbalances of mental stability, atherosclerosis, cystic fibrosis, diabetes, obesity, Crohn's disease, and altered circadian rhythmicity, arthritis, inflammatory reactions or disorders, psoriasis and other skin diseases, autoimmune diseases, allergies, hypertension, anxiety disorders, schizophrenia and other psychoses, osteoporosis, muscular dystrophy, amyotrophic lateral sclerosis or circadian rhythm-related conditions.

53. A method for identifying a population of low copy number CpG islands and GC rich regions, comprising:

cleaving genomic DNA with both a restriction enzyme that cleaves at a recognition site comprising adenosine and thymidine residues and a restriction endonuclease that cleaves at an unmethylated restriction site comprising cytidine and guanosine residues, to generate a population of restriction fragments excluding those that are methylated;

cloning restriction fragments of at least 200 nucleotides in length from the population of restriction fragments, in a restriction negative bacteria to generate a first library;

cleaving cloned DNA of the first library with a restriction enzyme that cleaves DNA at a restriction site within a CpG island or GC rich region;

excluding CpG island or GC rich region fragments that contain repetitive elements while leaving low copy CpG island or GC rich region fragments intact, thereby producing a population of low copy number CpG islands and GC rich regions.

54. The method of claim 53, further comprising cloning the restriction fragments containing low copy CpG islands and GC rich regions to form a library containing a plurality of low copy CpG island or GC rich region DNA.

55. The method of claim 53, wherein the excluding is by optionally cleaving cloned DNA of the first library with a restriction enzyme that cleaves DNA at a restriction site within a CpG island or GC rich region repeat sequence or using a methylated CpG binding column.

56. A library produced by the method of claim 53.

57. The method of claim 53, wherein the GC rich regions are CpG islands.

58. A method for identifying the DNA methylation status at a cytosine residue of a CpG sequence in a genomic DNA sample from the subject, wherein hypomethylation of CpG sequences compared to a methylated DNA control sample is indicative of a disease present within the subject.

59. A method according to claim 58, characterized in that a set of CpG positions comprises at least 3 CpG positions that are located in the regulatory region of the same gene.

60. A method according to claim 58 or **59**, characterized in that the methylation state of at least 3 different sets of CpG positions is identified.

61. The method of any of claims **1**, **6**, **31**, **34**, **48** or **53**, wherein the GC rich region is in an intron.

62. The method of any of claims **1**, **6**, **31**, **34**, **48** or **53**, wherein the GC rich region is in an exon.

63. The method of any of claims **1**, **6**, **31**, **34**, **48** or **53**, wherein the GC rich region is in a regulatory region.

* * * * *

US007078168B2

(12) **United States Patent**
Sylvan

(10) **Patent No.: US 7,078,168 B2**
(45) **Date of Patent: Jul. 18, 2006**

(54) **METHOD FOR DETERMINING ALLELE FREQUENCIES**

(75) Inventor: **Anna Sylvan**, Uppsala (SE)

(73) Assignee: **Biotage AB**, Uppsala (SE)

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 0 days.

(21) Appl. No.: **10/085,774**

(22) Filed: **Feb. 27, 2002**

(65) **Prior Publication Data**

US 2003/0082566 A1 May 1, 2003

Related U.S. Application Data

(60) Provisional application No. 60/271,703, filed on Feb. 27, 2001.

(51) **Int. Cl.**
C07H 21/04 (2006.01)
C12Q 1/68 (2006.01)

(52) **U.S. Cl.** **435/6**; 435/91.1; 435/91.2; 536/23.1; 536/24.3

(58) **Field of Classification Search** 435/6, 435/91.2, 91.1, 174; 536/23.1, 24.3
See application file for complete search history.

(56) **References Cited**

U.S. PATENT DOCUMENTS

6,020,137	A *	2/2000	Lapidus et al.	435/6
6,287,778	B1 *	9/2001	Huang et al.	435/6
6,566,101	B1 *	5/2003	Shuber et al.	435/91.2
6,750,018	B1 *	6/2004	Kambara et al.	435/6

FOREIGN PATENT DOCUMENTS

WO	WO 9828440	A1 *	7/1998

OTHER PUBLICATIONS

Nyren et al. "Detection of Single-Base changes using a bioluminometric primer extension assay." Analytical Biochemistry. Vo 244, pp. 367-373, 1997.*
Breen et al. "Determining SNP allele frequencies in DNA pools." BioTechniques. vol. 28, No. 3, pp. 464-470, Mar. 2000.*
Arnheim, N. et al. 1985. *Proc. Natl. Acad. Sci.* 82: 6970-6974.
Breen, G. et al. 1999. *Mol Cell Probes* 13(5): 359-365.
Breen, G. et al. 2000. *Biotechniques* 28(3): 464-3, 470.
Collins, H.E. et al. 2000. *Hum. Genet.* 106(2): 218-26.
Germer, S. et al. 2000. *Genome Res.* 10(2): 258-266.
Kruglyak, L. 1999. *Nat. Genet.* 22: 139-144.
Kwok, P. et al. 1994. *Genomics* 23: 138-144.
Pacek, P. et al. 1993. *PCR Methods Applic.* 2: 313-317.
Risch, N. and Merikangas, K. 1996. *Science* 273: 1516-1517.
Risch, N. and Teng, J. 1998. *Genome Res.* 8: 1273-1288.
Shaw, S. H. et al. 1998. *Genome Res.* 8: 111-123.
Bellman et al. 2000. Poster entitled "Reliable SNP assessment and allele frequency determination by Pyrosequencing" presented Apr. 9, 2000 at Human Genome Project Meeting, Vancouver.
Ekstrom. 2000. Printed copy of presentation slides and conference program for talk given Feb. 27, 2000 at Genomic Opportunities Conference, San Francisco.

* cited by examiner

Primary Examiner—Jeanine A. Goldberg
(74) *Attorney, Agent, or Firm*—Dorsey & Whitney LLP

(57) **ABSTRACT**

The present invention related to a method of determining the frequency of an allele in a population of nucleic acid molecules. The method comprises pooling the nucleic acid molecules of a population of nucleic acids, performing primer extension reactions using a primer which binds at a predetermined site located in nucleic acid molecules, and obtaining a pattern of nucleotide incorporation.

16 Claims, 16 Drawing Sheets

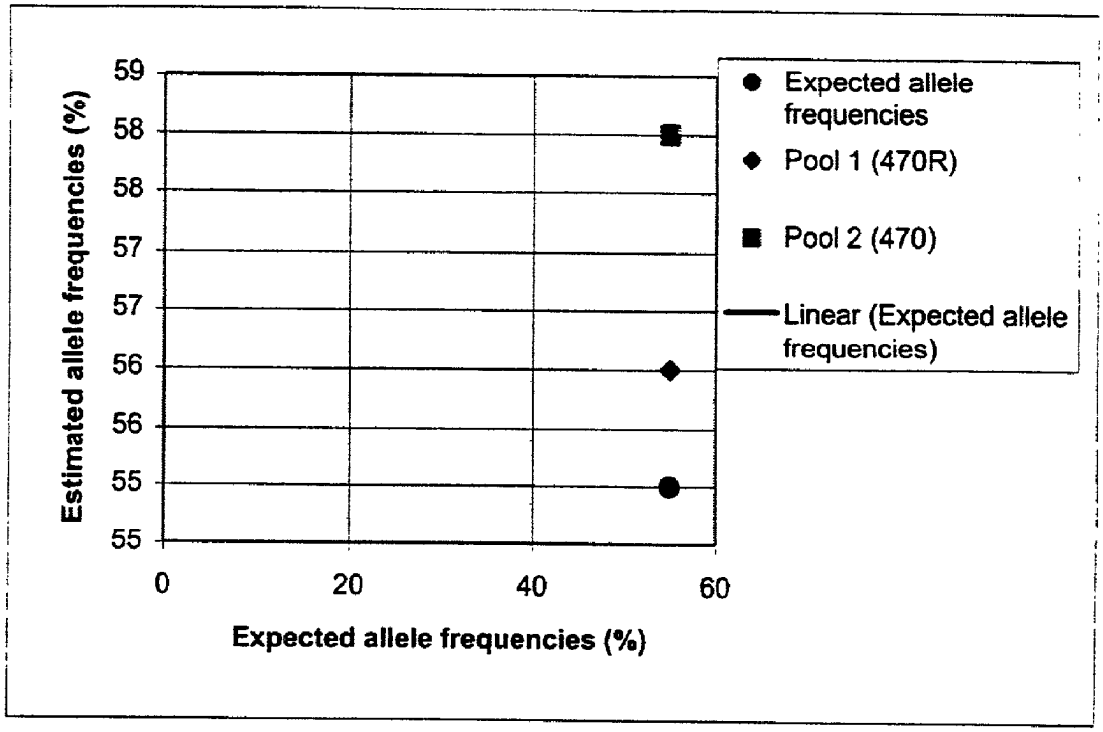


Figure 1 a

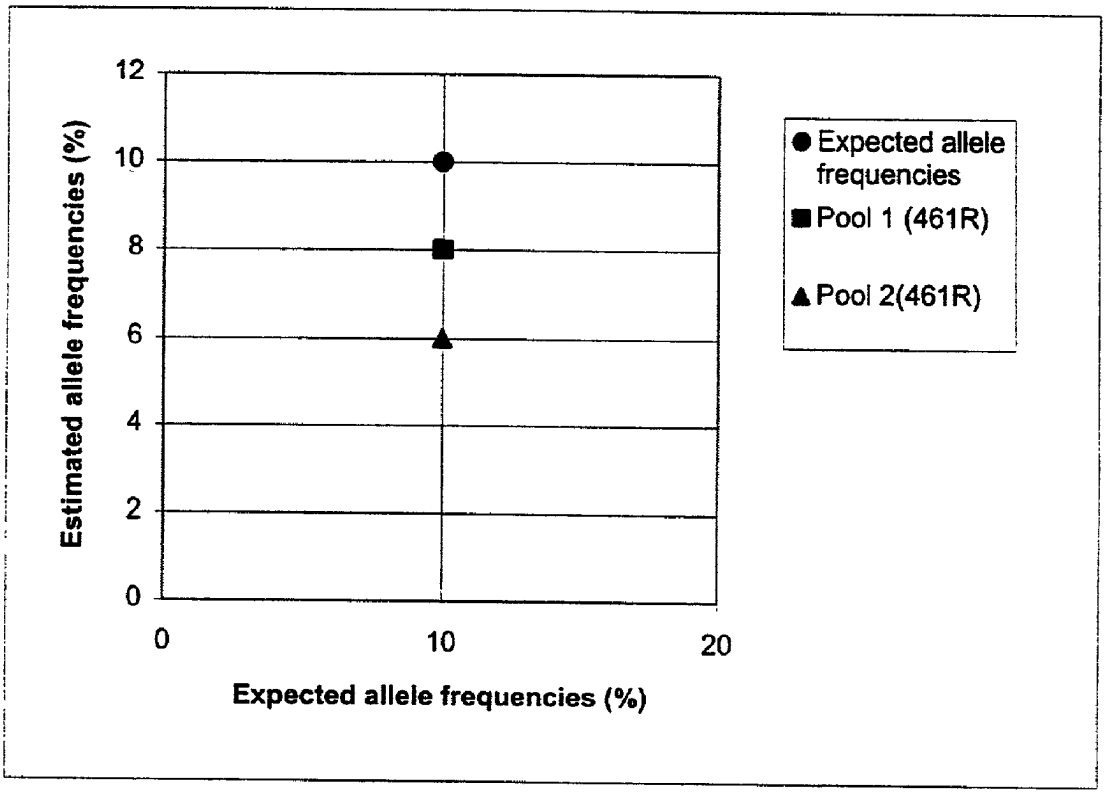


Figure 1 b

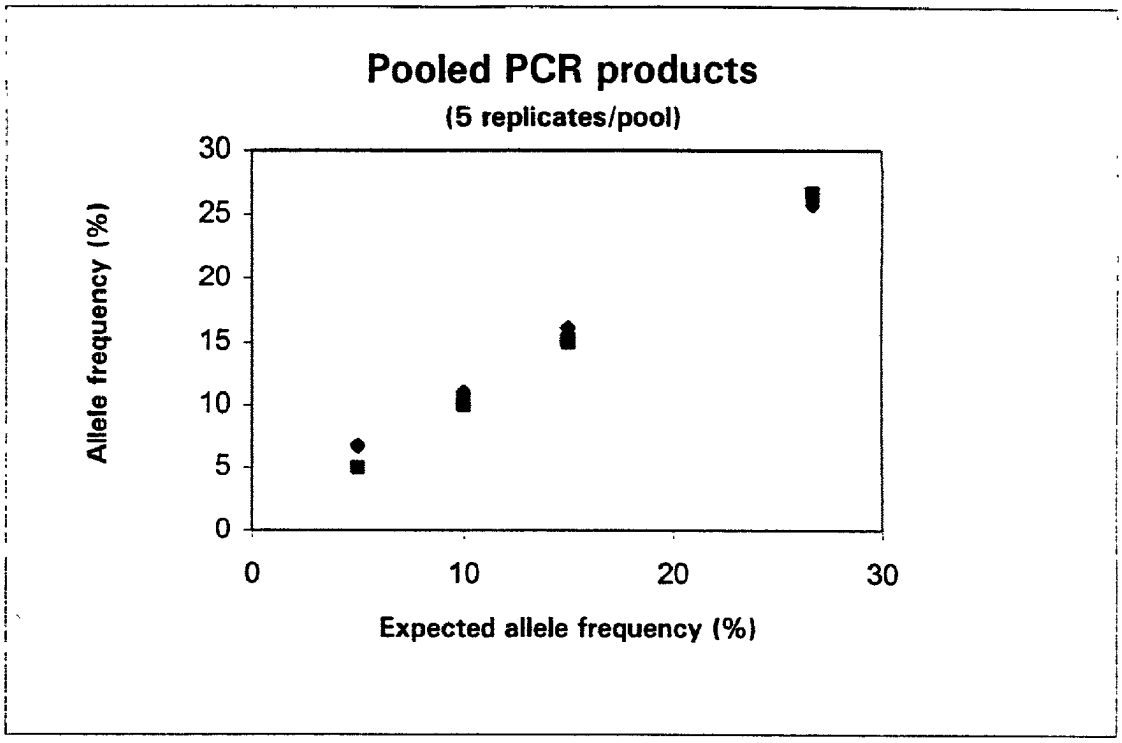

Pooled PCR products
(5 replicates/pool)
Allele frequency (%)
30
25
20
15
10
5
0
0
10
20
30
Expected allele frequency (%)

Fig. 2a

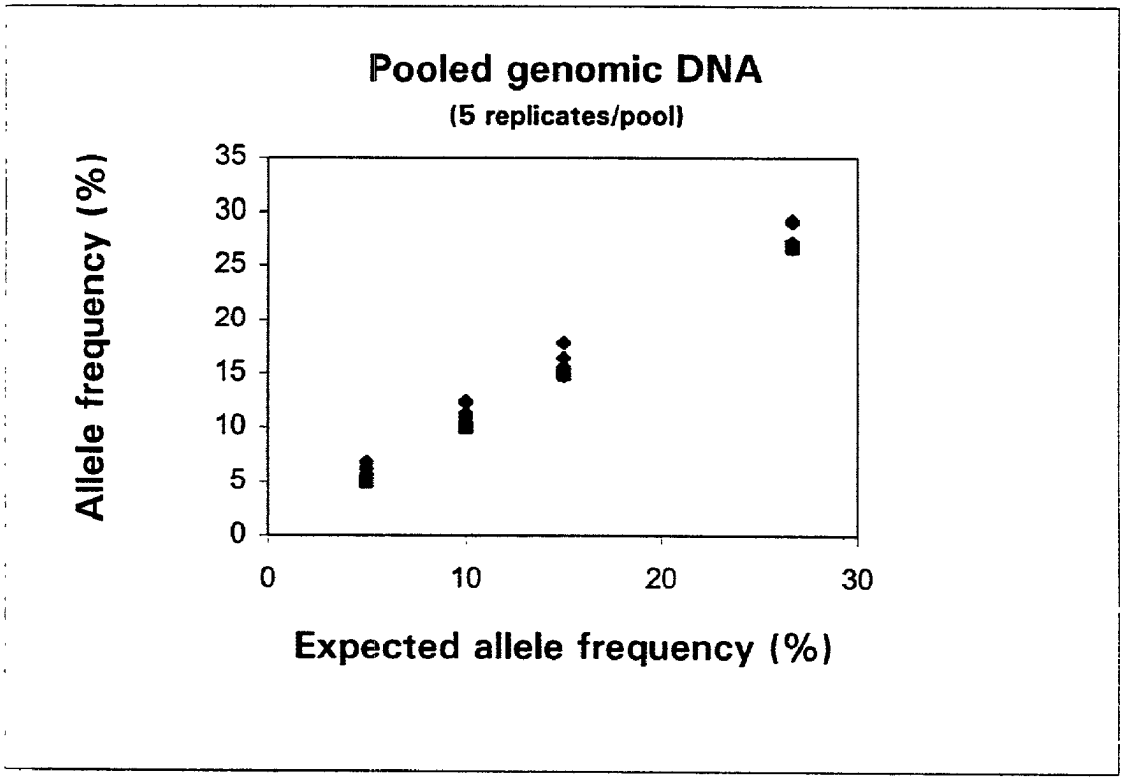

Pooled genomic DNA
(5 replicates/pool)
Allele frequency (%)
35
30
25
20
15
10
5
0
0
10
20
30
Expected allele frequency (%)

Fig. 2b

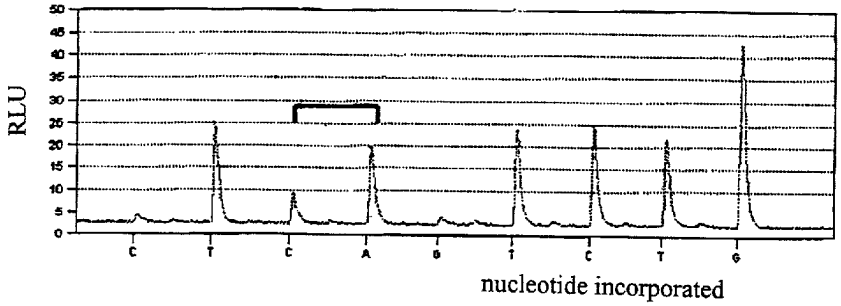


Fig. 3a

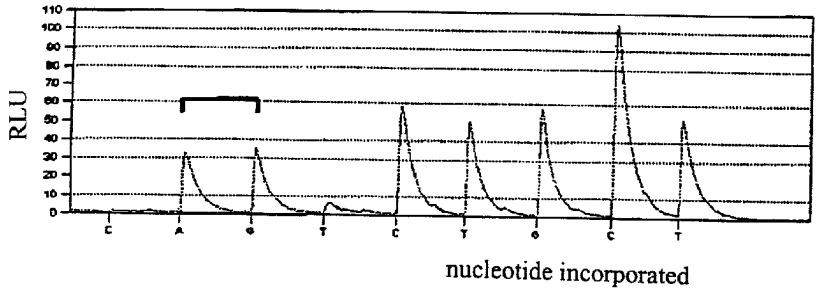


Fig. 3b

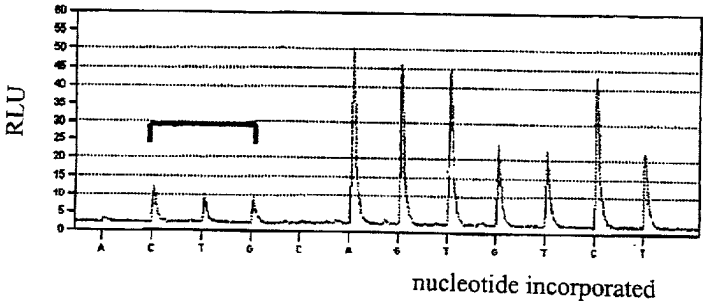


Fig. 3c

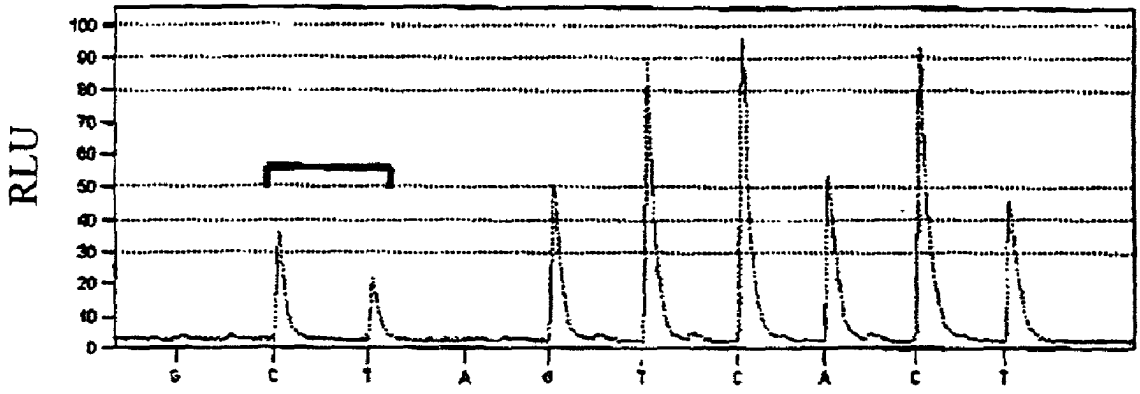

RLU
100
90
80
70
60
50
40
30
20
10
0
G C T A G T C A C T
nucleotide incorporated

Fig. 3d

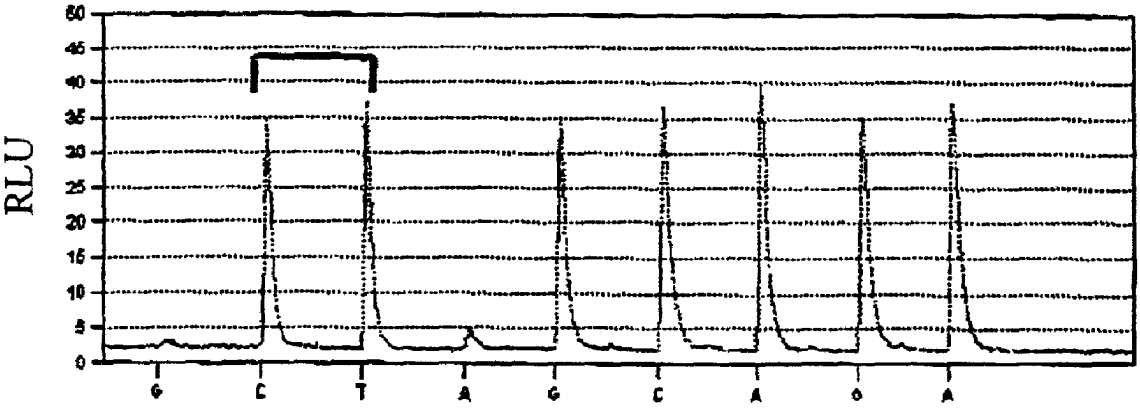

RLU
50
45
40
35
30
25
20
15
10
5
0
G C T A G C A G A
nucleotide incorporated

Fig. 3e

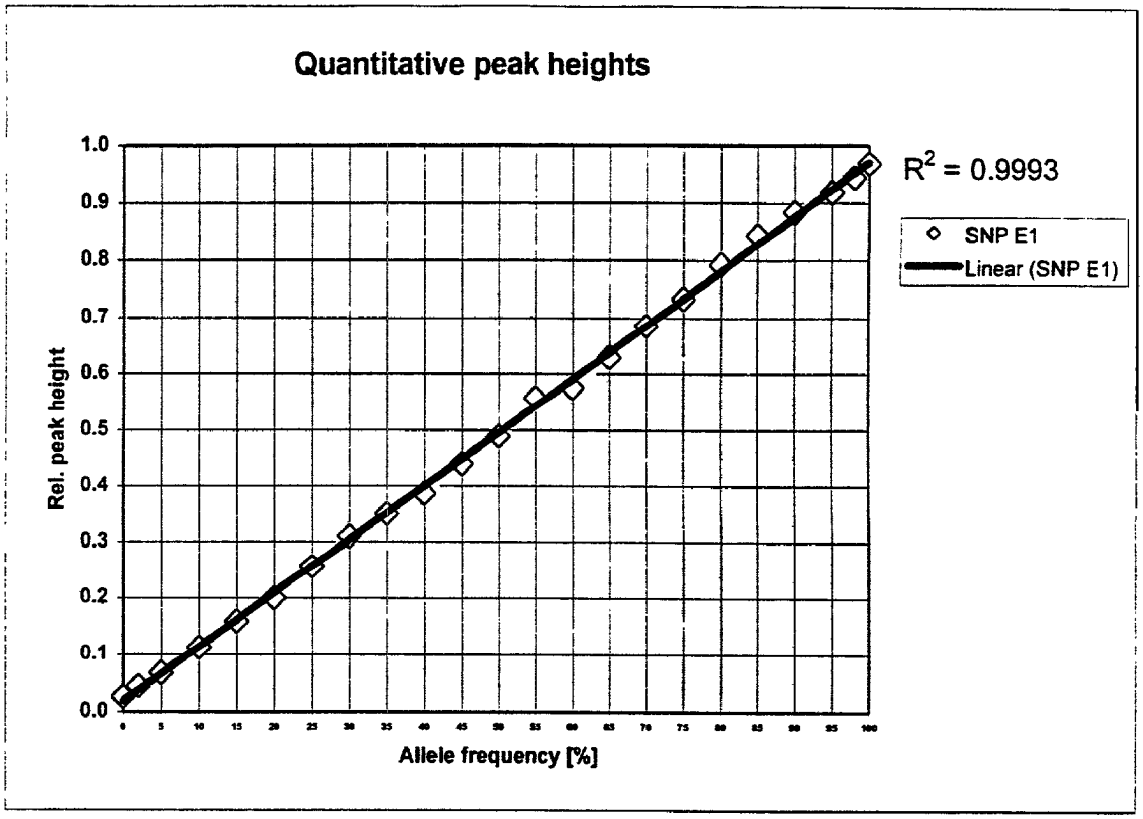

Quantitative peak heights
Rel. peak height
Allele frequency [%]
R^2 = 0.9993
SNP E1
Linear (SNP E1)

Fig. 4a

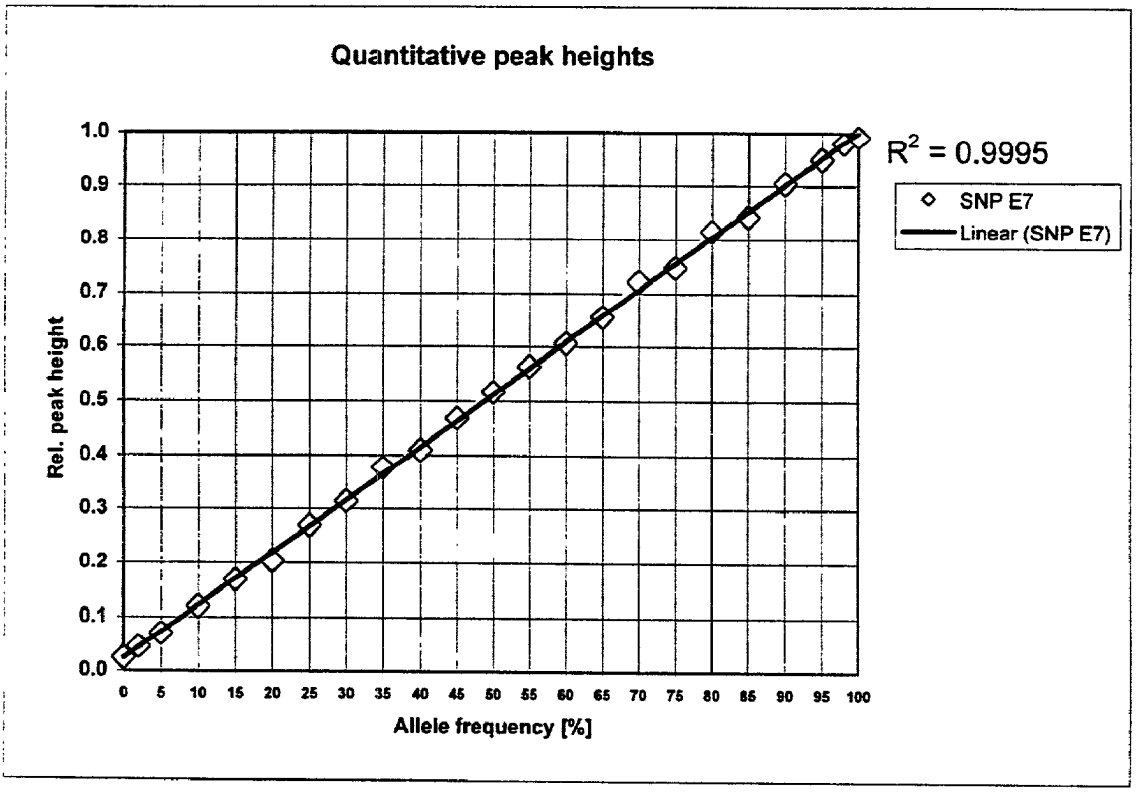

Quantitative peak heights
Rel. peak height
Allele frequency [%]
R^2 = 0.9995
SNP E7
Linear (SNP E7)

Fig. 4b

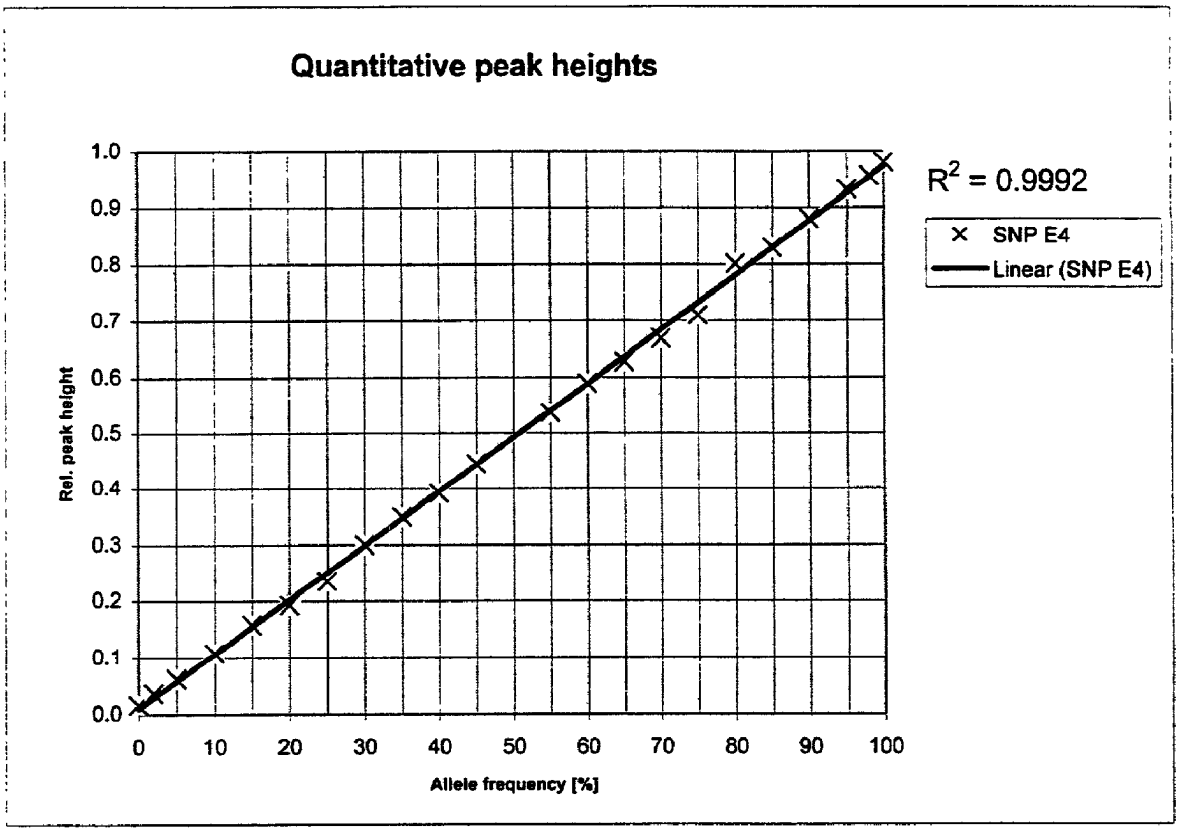

Quantitative peak heights
Rel. peak height
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
10
20
30
40
50
60
70
80
90
100
Allele frequency [%]
R2 = 0.9992
SNP E4
Linear (SNP E4)

Fig. 4c

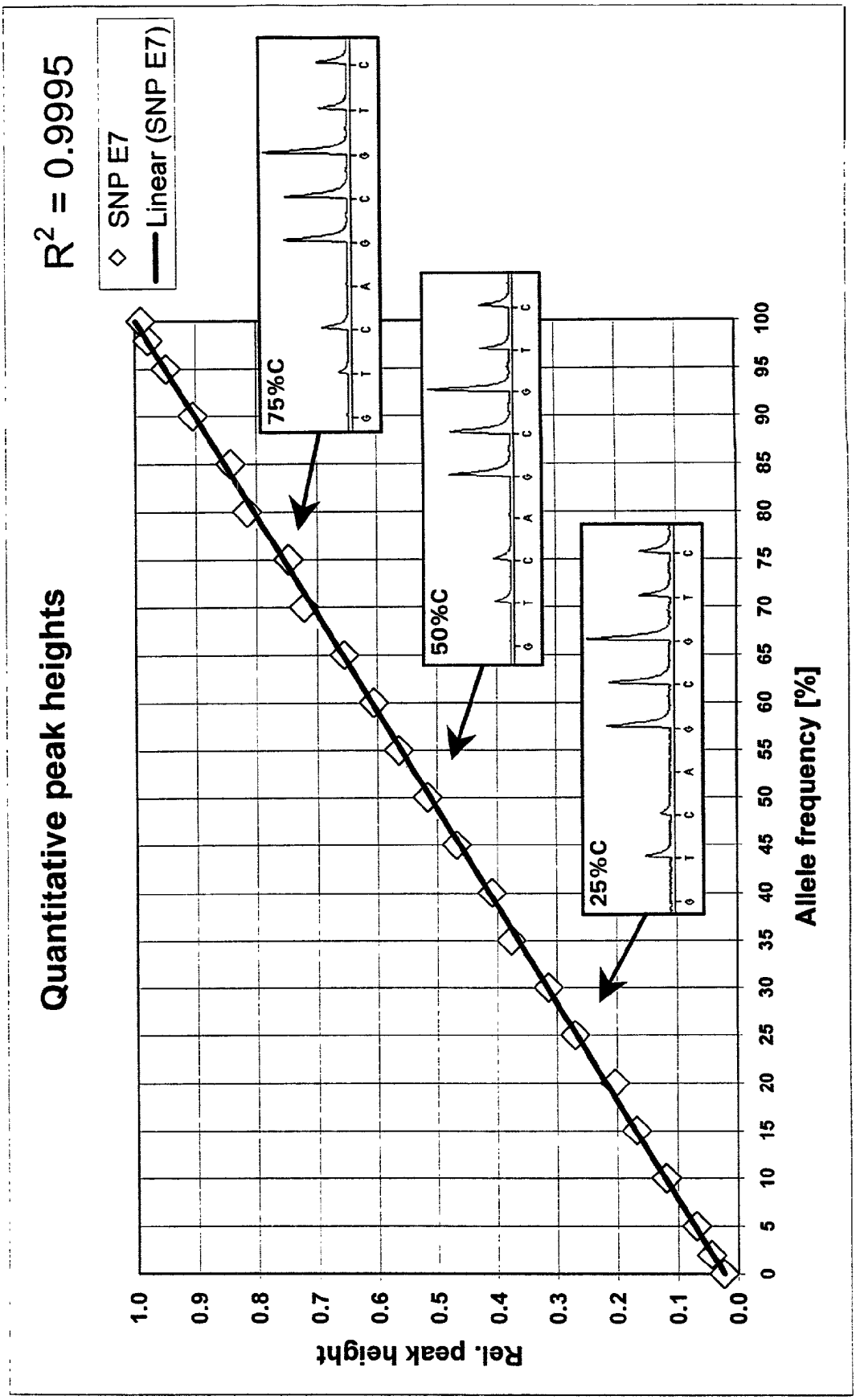

Quantitative peak heights
R^2 = 0.9995
SNP E7
Linear (SNP E7)
75%C
50%C
25%C
Rel. peak height
Allele frequency [%]
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
0
5
10
15
20
25
30
35
40
45
50
55
60
65
70
75
80
85
90
95
100

Fig. 5

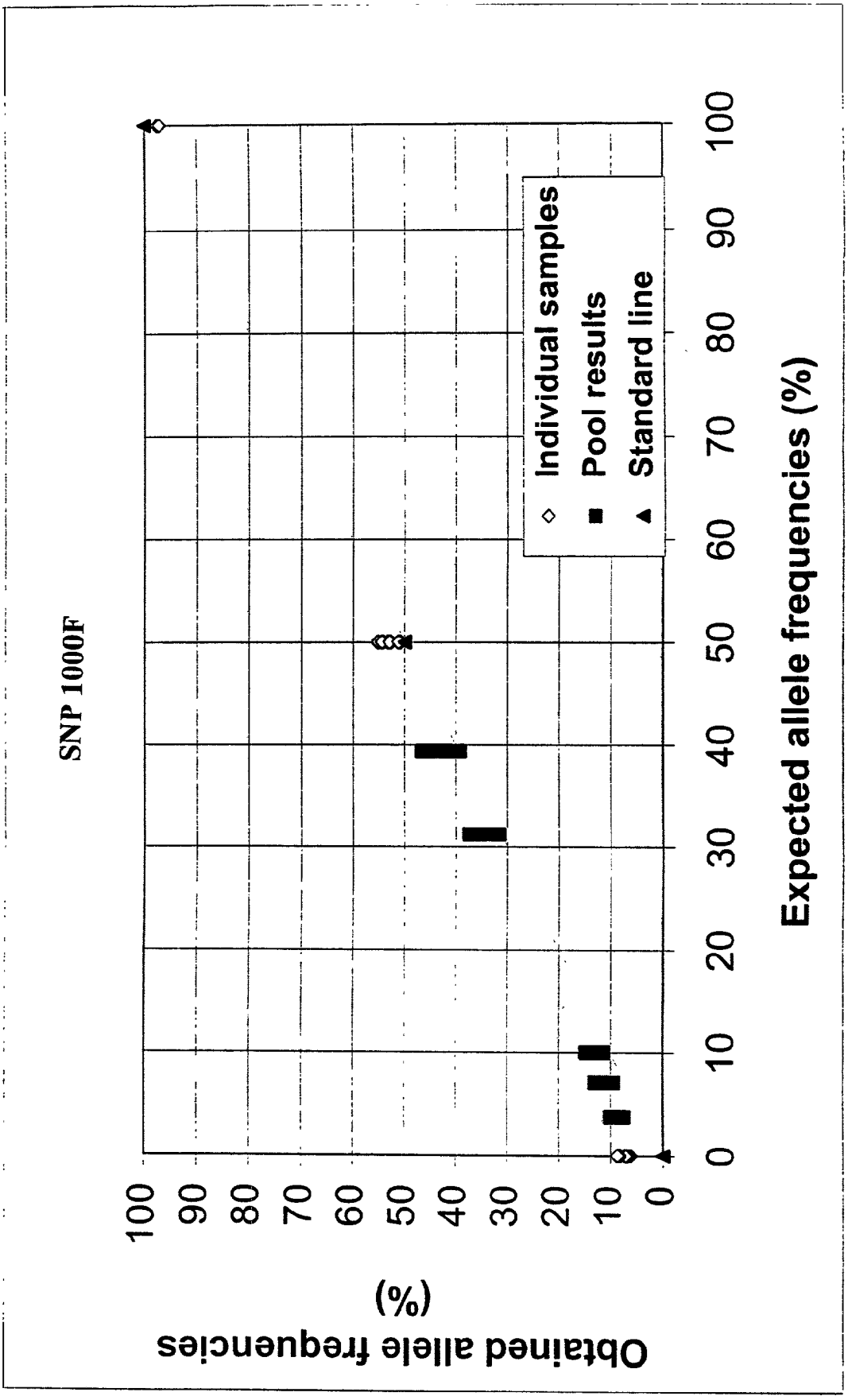

SNP 1000F
Individual samples
Pool results
Standard line
Expected allele frequencies (%)
Obtained allele frequencies (%)
0
10
20
30
40
50
60
70
80
90
100

Fig 6

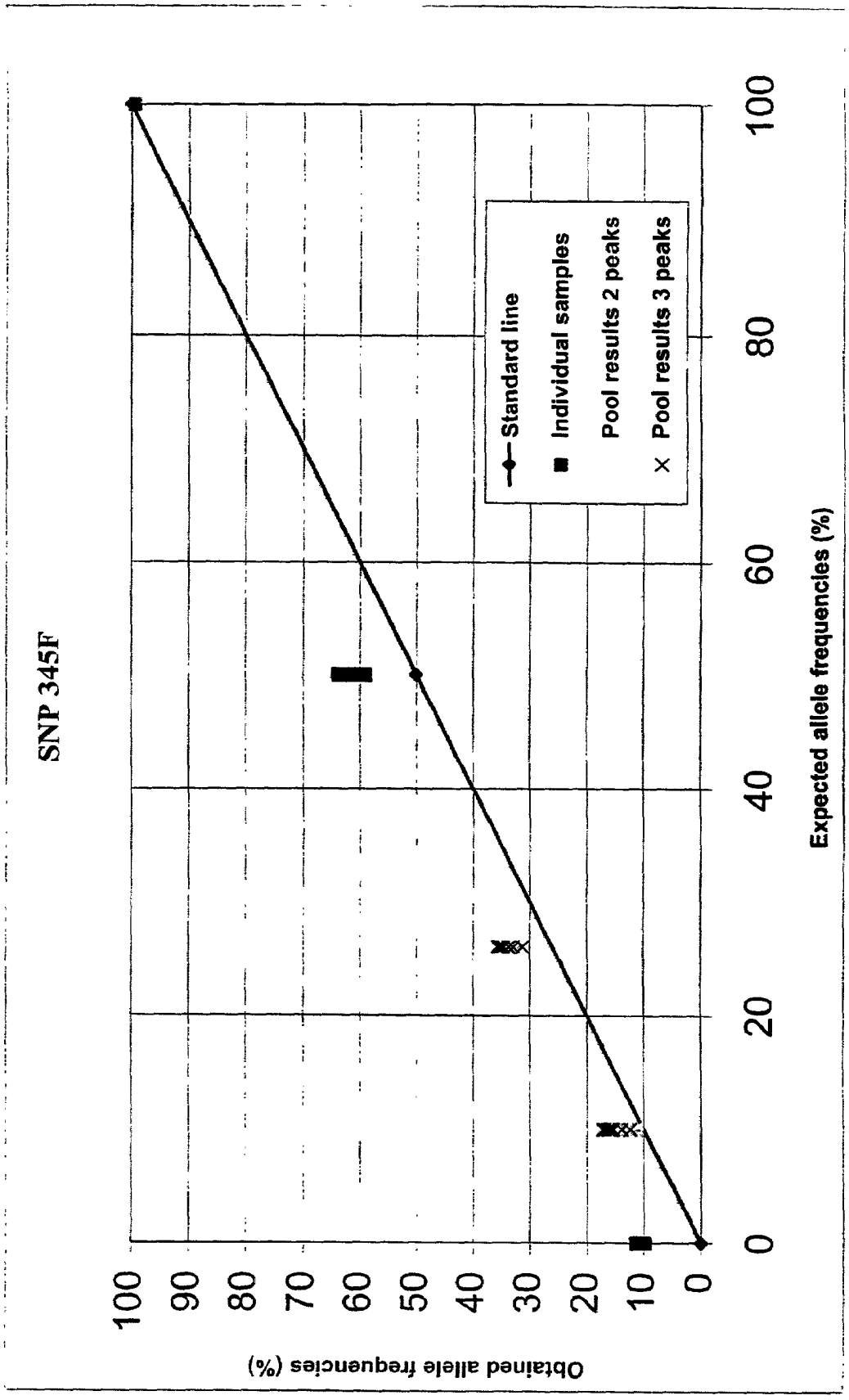

SNP 345F
Obtained allele frequencies (%)
100
90
80
70
60
50
40
30
20
10
0
Expected allele frequencies (%)
0
20
40
60
80
100
Standard line
Individual samples
Pool results 2 peaks
Pool results 3 peaks

Fig. 7

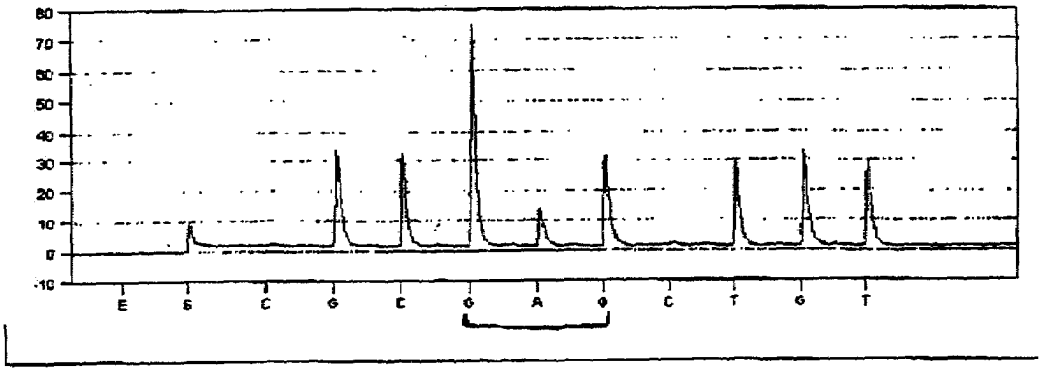

80
70
60
50
40
30
20
10
0
-10
E
S
C
G
C
G
A
G
C
T
G
T
nucleotide incorporated

Fig. 8a

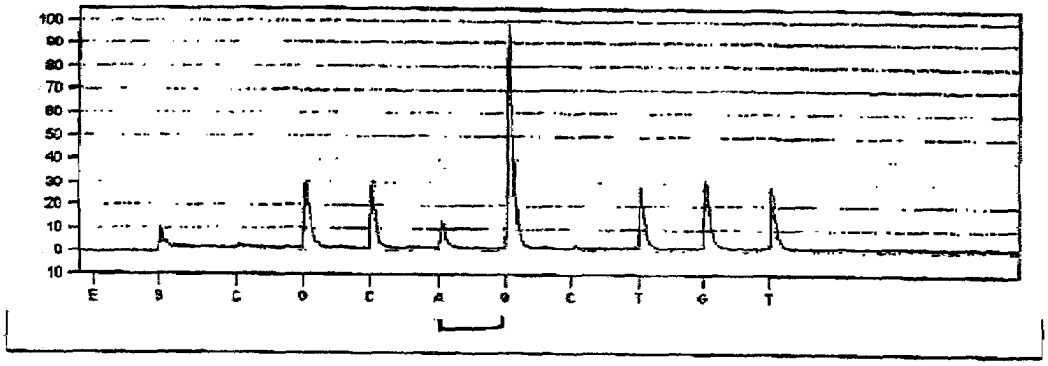

100
90
80
70
60
50
40
30
20
10
0
10
E
S
C
G
C
A
G
C
T
G
T
nucleotide incorporated

Fig. 8b

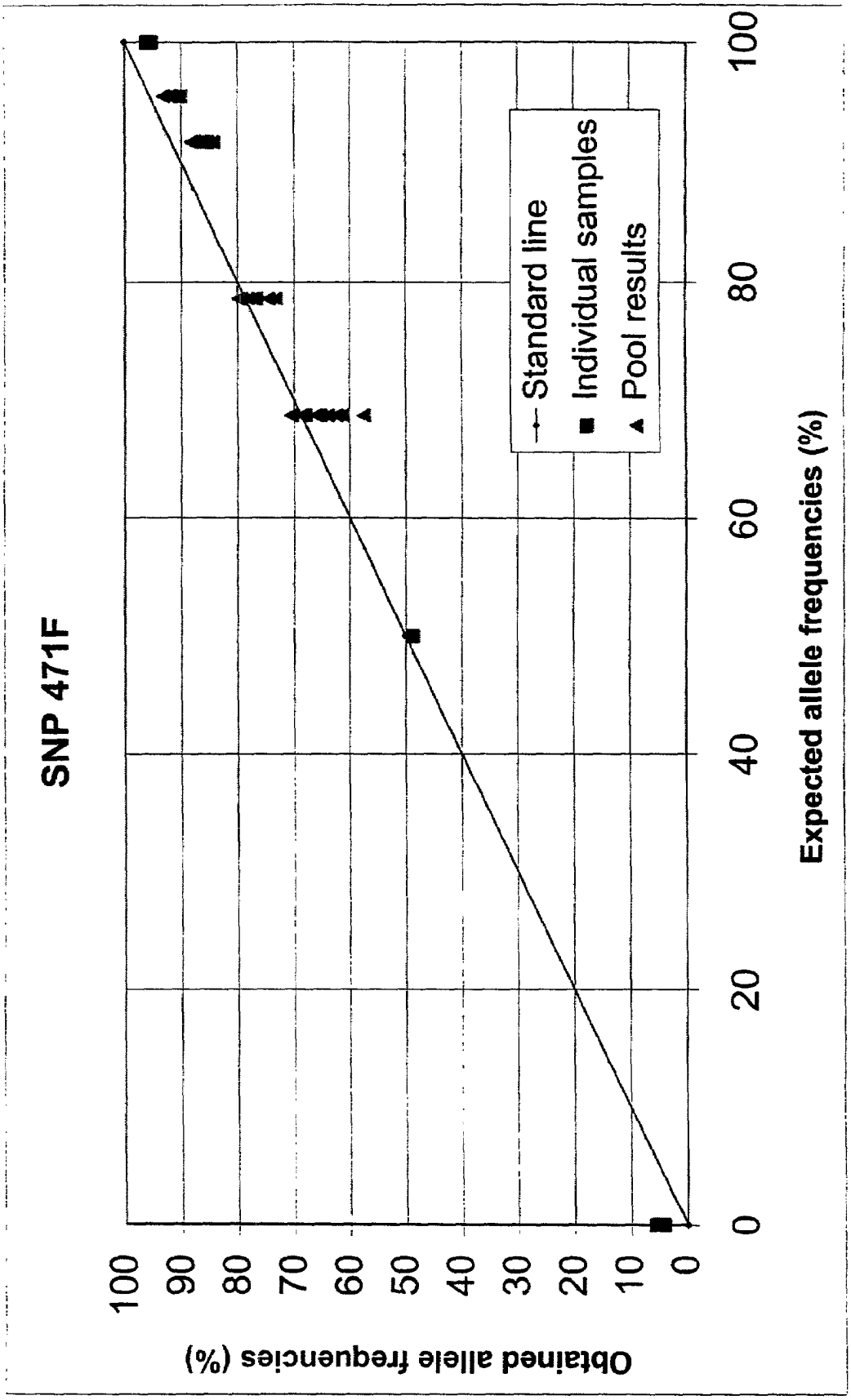

SNP 471F
Standard line
Individual samples
Pool results
Expected allele frequencies (%)
Obtained allele frequencies (%)
0
20
40
60
80
100
0
10
20
30
40
50
60
70
80
90
100

Fig. 9

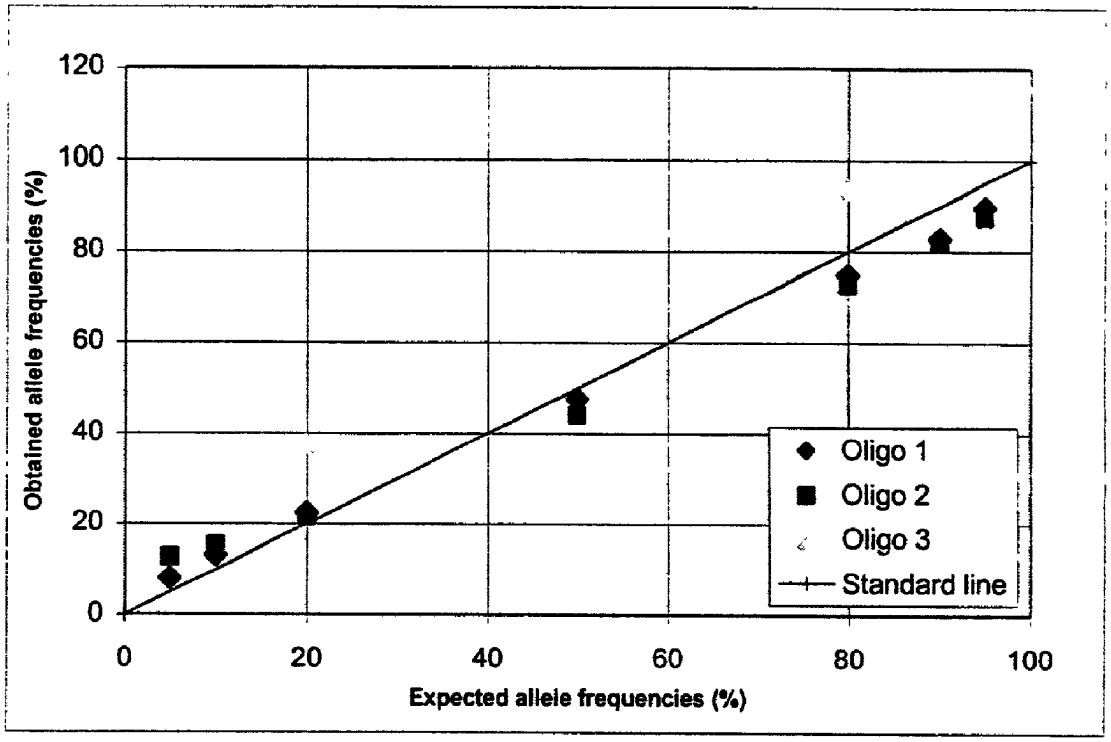


Fig. 10a

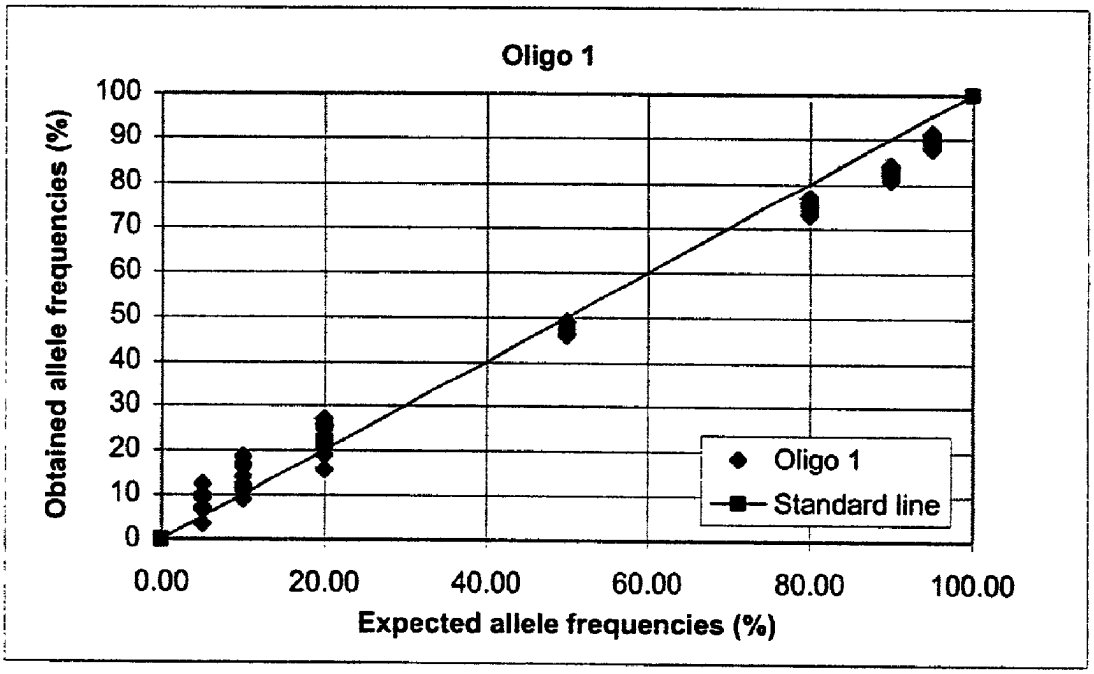


Fig. 10b

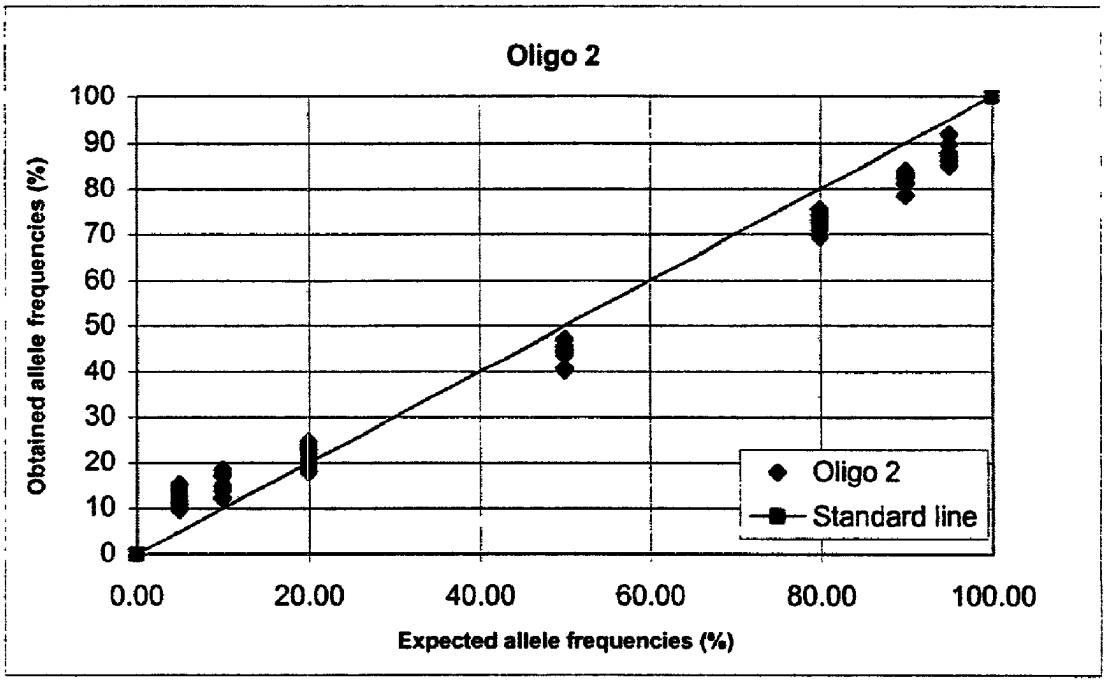


Fig. 10c

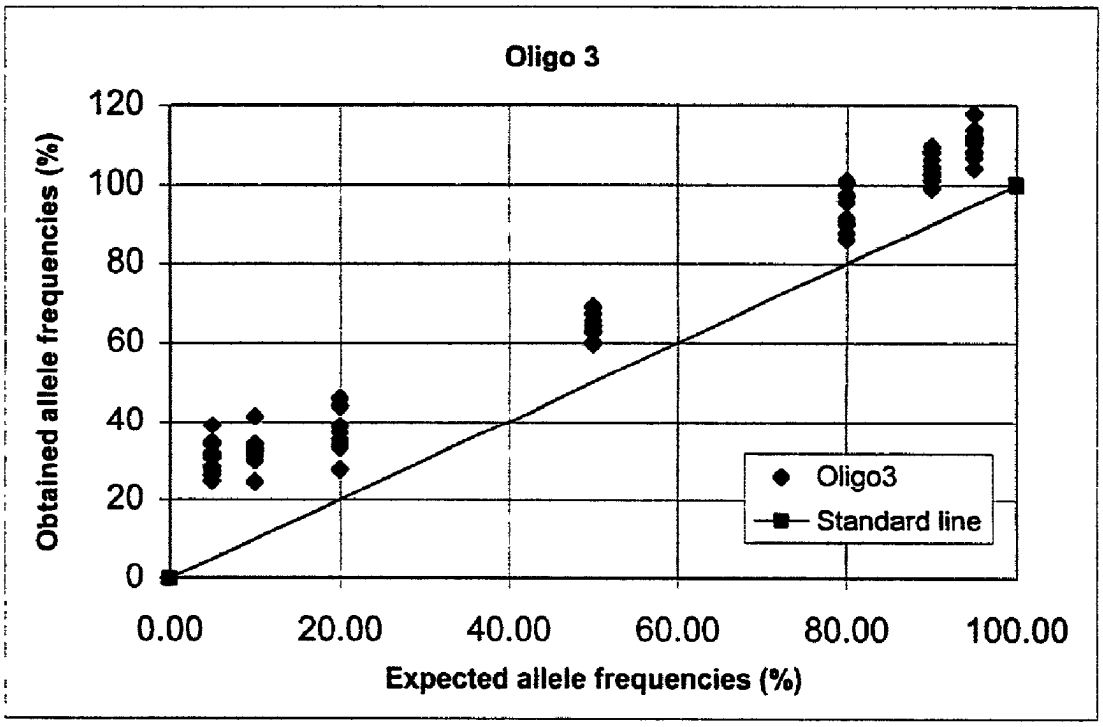


Fig. 10d

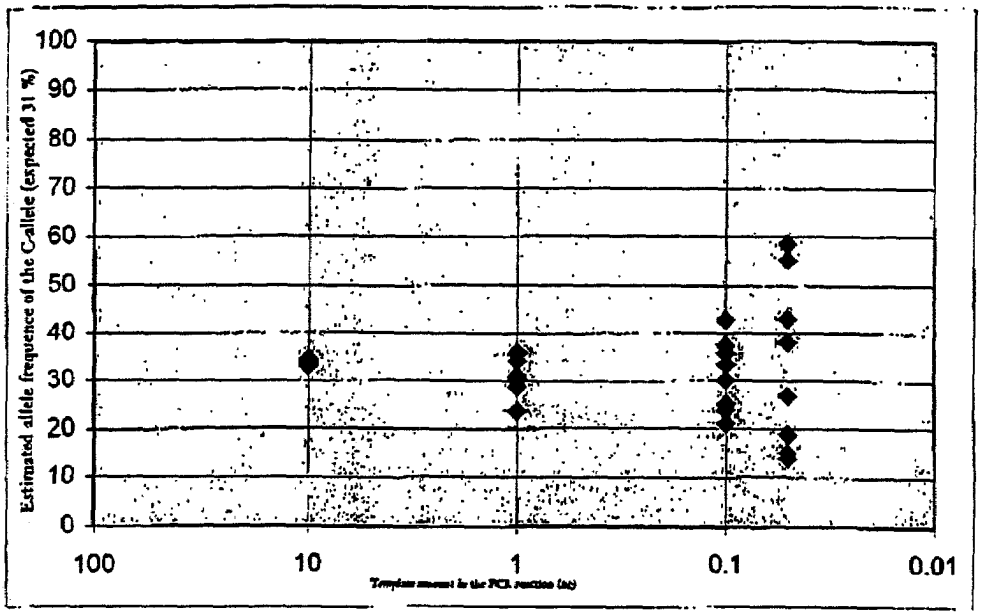


Fig. 11a

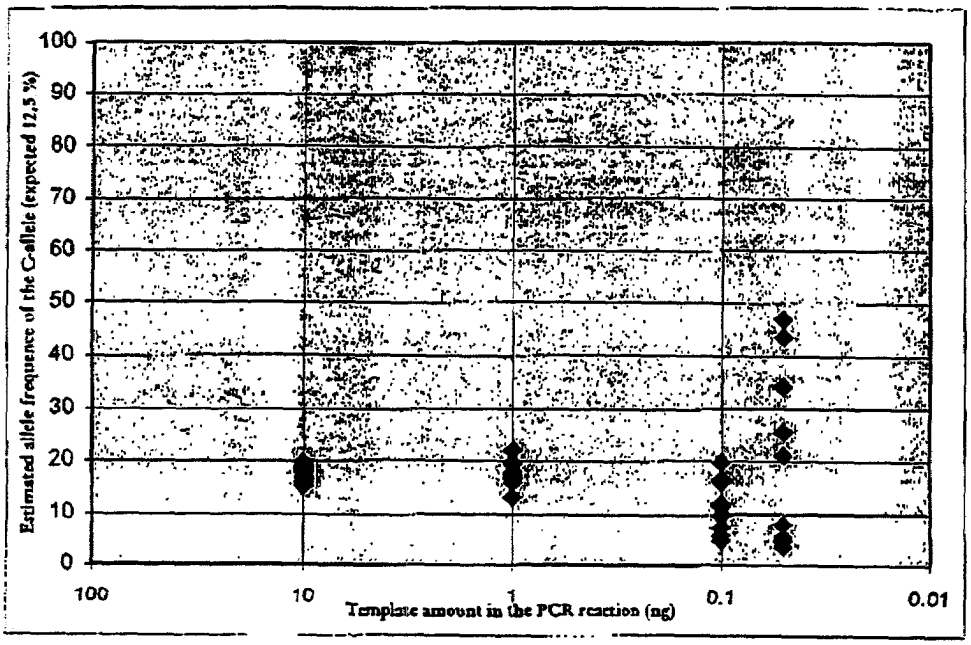


Fig. 11b

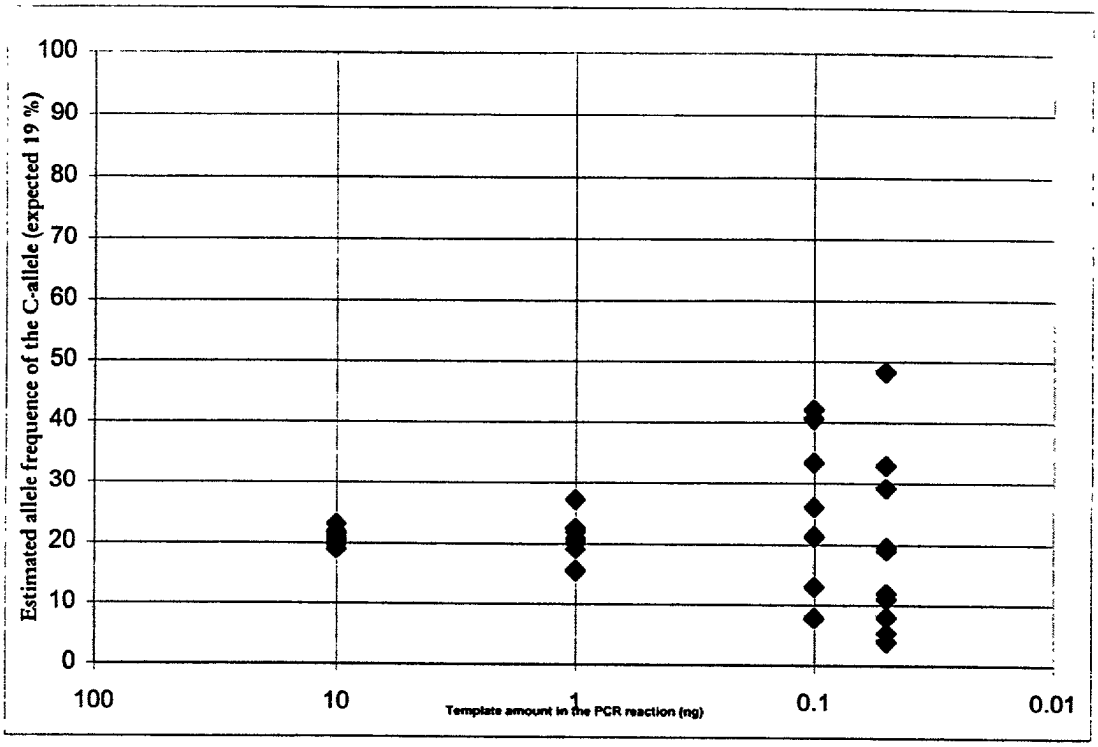


Fig. 11c

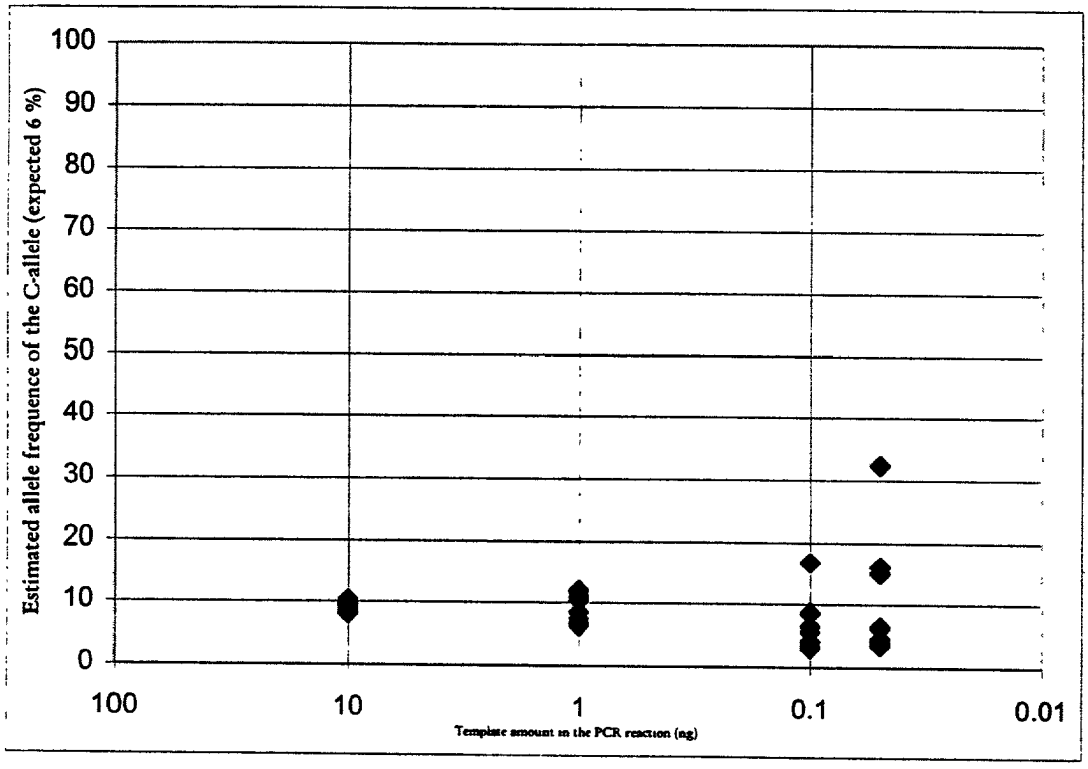


Fig. 11d

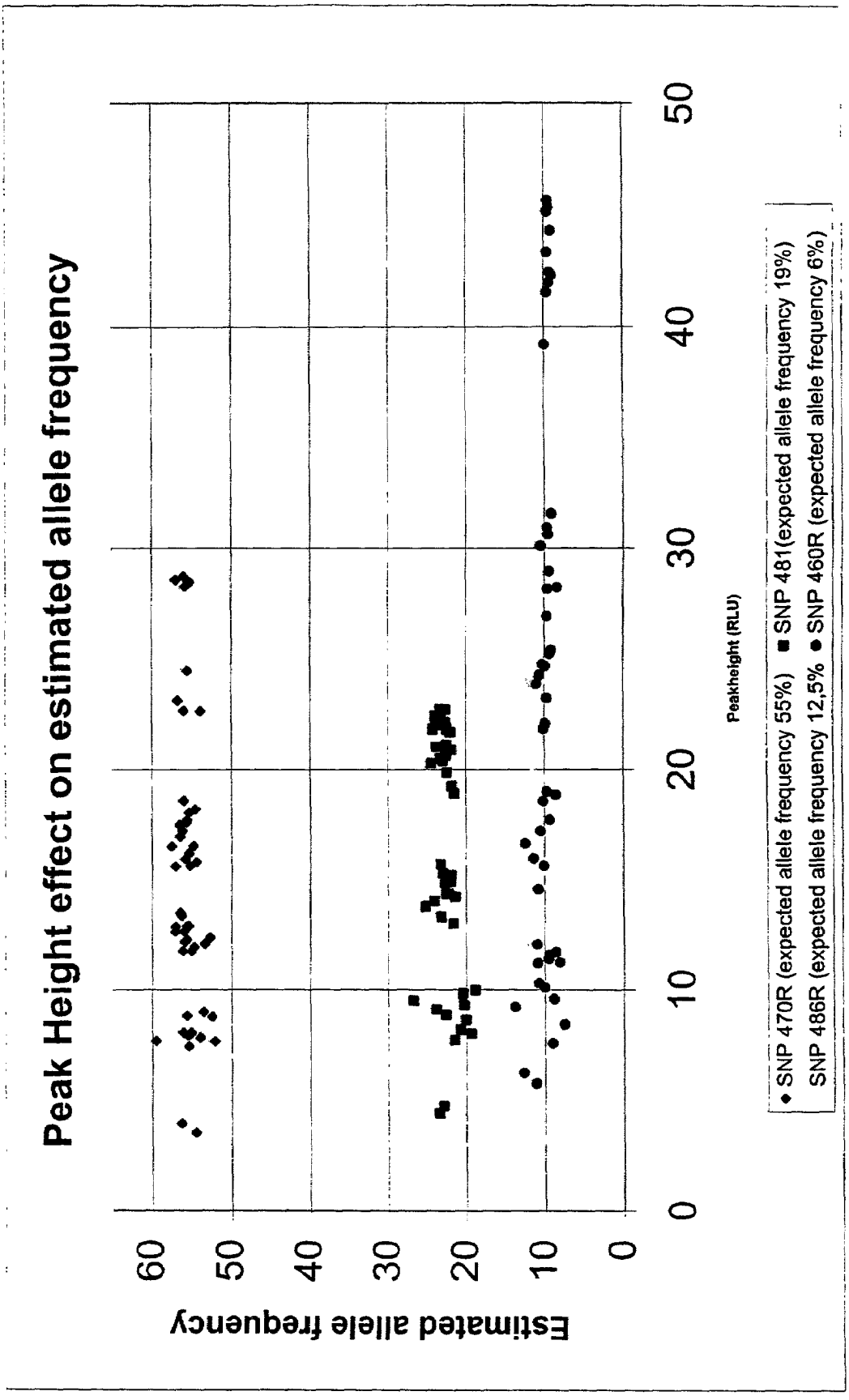

Peak Height effect on estimated allele frequency
Estimated allele frequency
0
10
20
30
40
50
60
0
10
20
30
40
50
Peakheight (RLU)
◆ SNP 470R (expected allele frequency 55%)
■ SNP 481(expected allele frequency 19%)
SNP 486R (expected allele frequency 12,5%
● SNP 460R (expected allele frequency 6%)

Fig. 12

METHOD FOR DETERMINING ALLELE FREQUENCIES

CROSS-REFERENCE TO RELATED APPLICATION

This application claims the benefit of U.S. application Ser. No. 60/271,703 filed Feb. 27, 2001, the disclosure of which is incorporated herein by reference.

BACKGROUND OF THE INVENTION

The invention relates to a method of determining the frequency of an allele within a given population or group, and in particular to a method of determining allele frequencies for single nucleotide polymorphisms (SNPs) or other mutations or genetic variations (e.g. nucleotide insertions, additions or deletions, gene, chromosome or genome duplications (or multiplications) etc. in pooled nucleic acid samples or other samples (including single samples) which may contain allelic variants.

Individuals in populations will have genetic differences. The genetic differences may be represented as the individuals in the population having different alleles at a given locus. Alternatively genetic differences can be related to gene, chromosome, or whole genome duplications (or other multiplications). The allele frequency describes the fraction of the population exhibiting a particular allele. Over a whole population, there may be many different alleles at a particular locus. However, where the genetic difference occurs as alterations of a single nucleotide (single nucleotide polymorphisms or SNPs), generally only 2 alleles are present in the population, although triallelic or tetrallelic SNPs are known. Studies of allelic association in populations are one of the most useful and powerful methods for mapping genes/mutations that contribute to disease. Such studies require the determination of the genotype (i.e. which allele is present) at one or several loci in a population. The frequency of a particular allele in a given population can be assessed, and the association of that allele with a disease or other clinical condition (e.g. predisposition to disease, therapeutic responsibility etc.) can be studied.

Single nucleotide polymorphisms (SNPs) are regularly used for genetic association studies, and consist of single nucleotide substitutions. SNPs are normally biallelic markers (i.e. there are 2 alleles present in the population), and are the markers of choice for various types of genetic analysis, because of their high frequency in the genome. SNPs are found approximately once every 100 to 1000 bases in the human genome. An SNP has a prevalence of at least 1% in a given population. Further, they are stable, having much lower mutation rates than repeat sequences, for example. The analysis of SNPs is of great importance in several disciplines within the applied genomic field. Importantly, the nucleotide sequence variations that are most likely to be responsible for the functional changes of interest will be SNPs. Such variations are therefore of great interest, and many studies directed to identify functional SNPs contributing to (or associated with) a particular trait or disease ("phenotype") have been performed. Thus many diseases and conditions may be associated with (or linked to) single nucleotide polymorphisms, either alone or in combination. For example, in WO 00/22166, it has been suggested that a combination of SNPs within several genes gives a polymorphic pattern which may be used to predict the likelihood of developing cardiovascular disease. Obtaining reliable and accurate data on the frequencies of a given SNP allele in a given population without testing each member of the population would have a revolutionary impact on the efficiency and cost of analysis for large population studies.

However, the frequency of other genetic mutations or variants, e.g. insertion/addition/deletion mutations and gene, chromosome or genome duplications (in the sense of any number of multiplications or repeats), and those studied in cancer genetics and chromosomal abnormality (e.g. trisomy) cases, can be analysed by the method of the invention.

Allelic association means that across a given population, individuals who have a certain allele at one locus may have a statistically higher chance of developing a particular disease, for example. Thus, the possession of a particular allele can cause direct susceptibility to a disease. Alternatively, the possession of a particular allele may be indirectly linked to disease susceptibility via association with the "disease" allele.

Association studies attempt to find genes that influence or increase susceptibility to disease or traits in any organism. This involves determining the frequency of an allele from a population of organisms with that trait or disease and comparing the results with a control population that do not exhibit the disease or trait. Various statistical/mathematical methods are known and described in the art for assessing allele frequencies based on such studies. In order to perform large-scale association studies for single nucleotide polymorphisms, methods have included labourious and expensive individual genotyping of individual nucleic acid samples. Pooling of nucleic acid samples in order to obtain allele frequency information has been used to reduce the burden of genotyping individual samples. To date, most pooling investigations have centred on the use of microsatellite polymorphisms, with few methods developed for the rapid assessment of SNPs in a given population.

Studies on allele frequencies tend to rely on radiation-based methods, or gel electrophoresis, which have well-known drawbacks. A method of determining SNP allele frequency using allele-specific fluorescent probes in the Taqman® assay (Breen et al., Biotechniques 2000, 28(3) 464–470) has been developed by PE Biosystems. In this technique Taqman® probes are used to detect specific sequences in Polymerase Chain Reaction (PCR) products by employing the 5' 3' exonuclease activity of Taq polymerase. The Taqman® probe anneals to the target sequence between the traditional forward and reverse PCR primers. The Taqman® probe is labelled with a reporter fluorophore and a quencher fluorochrome. This technique relies on the possibility of designing allele specific probes that match the annealing temperature of the PCR primers. Moreover, the allele specificity of the probe is, in the case of SNPs, determined by one out of 17–30 bases. These restrictions make it hard to design allele specific probes showing good enough temperature discrimination not to bind to the other allele. Hence, the signal from such an assay might not always accurately represent the frequency of the probe specific allele. A disadvantage of this method may be in finding assay conditions where a mismatch results in clearly distinguishable difference in cleavage of the reporter fluorophore on the two alleles. Further, Taqman® probes have different dyes at the 5' and 3' ends and are therefore costly to produce, and must be carefully designed. Taqman requires two reactions in order to measure allele frequency, using a different probe in each of the two reactions, complementary to either allele. It would therefore be advantageous to develop a method of determining SNP allele frequencies in pooled nucleic acid in one reaction which was accurate,

reliable and that avoided the need for labels or relied on probe binding to the SNP site.

BRIEF SUMMARY OF THE INVENTION

It has now been found that a simple, reliable, reproducible and accurate method for determining the frequency of an allele in a given population, may be performed by pooling the nucleic acid sequences of the said population and performing a "primer-extension" type reaction, using primers designed for particular SNPs/alleles, and detecting the pattern of incorporation of nucleotides in said "primer-extension" reaction. The pattern may then be analysed to determine the frequency of each allele in the pooled nucleic acid.

The method is particularly suited to automation e.g. in systems where reaction and reagent dispensing steps take place in a microtitre plate format. The methods are particularly suited for finding SNP markers that are correlated to a certain trait, for example a specific disease, but may also find application in other allele frequency applications, such as SNP confirmation or analysis of mutations associated with cancer or chromosome abnormalities, especially abnormalities of chromosome number, and other mutations or variations involving duplication or loss of chromosomes or genes.

As described further below the present invention is advantageously based on a method of "sequencing-by-synthesis" (see e.g. U.S. Pat. No. 4,863,849 of Melamede). This is a term used in the art to define sequencing methods which rely on the detection of nucleotide incorporation during a primer-directed polymerase extension reaction. The four different nucleotides (i.e. A, G, T or C nucleotides) are added cyclically or sequentially (conveniently in known order), and the event of incorporation can be detected directly or indirectly. This detection reveals which nucleotide has been incorporated, and hence sequence information, when the nucleotide (base) which forms a pair (according to the normal rules of base pairing, A-T and C-G) with the next base in the template sequence is added, it will be incorporated into the growing complementary strand (i.e. the extended primer) by the polymerase, and this incorporation will trigger a detectable signal, the nature of which depends upon the detection strategy selected.

BRIEF DESCRIPTION OF THE DRAWINGS

FIG. **1***a* depicts the expected allele frequency (SNP 470R) and calculated allele frequency determined (estimated) via Pyrosequencing™. The results are plotted as estimated allele frequency versus expected allele frequency. Pool 1 has been calibrated according to Example 3, whereas the DNA concentration in pool 2 has been assayed via absorbance of light at 260 nm.

FIG. **1***b* depicts the expected allele frequency (SNP 461R) and calculated allele frequency determined (estimated) via Pyrosequencing™. The results are plotted as estimated allele frequency versus expected allele frequency. Pool 1 has been calibrated according to Example 3, whereas the DNA concentration in pool 2 has been assayed via absorbance of light at 260 nm. It should be noted that SNP 461R consistently gives a peak that is 3% too high, and the results shown are consistent with this.

FIG. **2***a* depicts the calculated allele frequency results of 4 pools of PCR products determined via Pyrosequencing™. 5 replicate reactions were performed on each pool. The results are plotted as estimated allele frequency versus expected allele frequency, both in percentage (%). The pools contained 27% G, 15% G, 10% G and 5% G. The calculated allele frequency value (shown as diamonds) are in close correlation to the expected values (shown as squares).

FIG. **2***b* depicts the calculated allele frequency results of 4 pools of genomic DNA samples determined via Pyrosequencing™. 5 replicate reactions were performed on each pool. The results are plotted as estimated allele frequency versus expected allele frequency, both in percentage (%). The pools contained 27% G, 15% G, 10% G and 5% G. The calculated allele frequency value (shown as diamonds) are in close correlation to the expected values (shown as squares).

FIG. **3***a* shows DNA sequencing on pooled genomic DNA over SNP 470R, the expected sequence of which is T[C/A] TCTGG. 40 μl PCR product was incubated with 15 μl magnetic beads (10 μg/μl) and 25 μl 2×BW buffer. Pyrosequencing™ was then performed on a PSQ™ 96 system instrument using Pyrosequencing™ SNP reagent kit. The peak heights were measured in order to calculate the frequency of the allele. The results are shown generally as nucleotide incorporated (i.e. A, C, G or T) versus amount of light released (in RLU). The 2 nucleotide incorporations which relate to the SNP are marked. The experimental conditions are as described in Example 4.

FIG. **3***b* shows DNA sequencing on pooled genomic DNA over SNP EU4, the expected sequence of which is [A/G] CTGCCT. 40 μl PCR product was incubated with 15 μl magnetic beads (10 μg/μl) and 25 μl 2×BW buffer. Pyrosequencing™ was then performed on a PSQ™ 96 system instrument using Pyrosequencing™ SNP reagent kit. The peak heights were measured in order to calculate the frequency of the allele. The results are shown generally as nucleotide incorporated (i.e. A, C, G or T) versus amount of light released (in RLU). The 2 nucleotide incorporations which relate to the SNP are marked. The experimental conditions are as described in Example 4.

FIG. **3***c* shows DNA sequencing on pooled genomic DNA, over SNP 466F, the sequence of the nucleic acid should be [C/T/G]AAGGTTGTCCT (SEQ. ID NO: 1) 40 μl PCR product was incubated with 15 μl magnetic beads (10μg/μl) and 25 μl 2×BW buffer. Pyrosequencing™ was then performed on a PSQ™ 96 system instrument using Pyrosequencing™ SNP reagent kit. The peak heights were measured in order to calculate the frequency of the allele. The results are shown generally as nucleotide incorporated (i.e. A, C, G or T) versus amount of light released (in RLU). The 3 nucleotide incorporations which relate to the SNP are marked. The experimental conditions are as described in Example 4.

FIG. **3***d* shows DNA sequencing on pooled genomic DNA, over SNP 465R, the sequence of the nucleic acid should be [CIT]GTTCCACCT (SEQ. ID NO: 2). 40 μl PCR product was incubated with 15 μl magnetic beads (10 μg/μl) and 25 μl 2×BW buffer. Pyrosequencing™ was then performed on a PSQ™ 96 system instrument using Pyrosequencing™ SNP reagent kit. The peak heights were measured in order to calculate the frequency of the allele. The results are shown generally as nucleotide incorporated (i.e. A, C, G or T) versus amount of light released (in RLU). The 2 nucleotide incorporations which relate to the SNP are marked. The experimental conditions are as described in Example 4.

FIG. **3***e* shows DNA sequencing on pooled genomic DNA, over SNP 461R, the sequence of the nucleic acid should be [C/T]TGCAGA. 40 μl PCR product was incubated with 15 μl magnetic beads (10 μg/μl) and 25 μl 2×BW buffer. Pyrosequencing™ was then performed on a PSQ™

96 system instrument using Pyrosequencing™ SNP reagent kit. The peak heights were measured in order to calculate the frequency of the allele. The results are shown generally as nucleotide incorporated (i.e. A, C, G or T) versus amount of light released (in RLU). The 2 nucleotide incorporations which relate to the SNP are marked. The experimental conditions are as described in Example 4.

FIG. **4***a* depicts graphically relative peak heights from a Pyrosequencing reaction plotted against allele frequency. The SNP analysed was SNPE1. 5 pmol pooled DNA PCR product was incubated with 17.5 μl magnetic beads, and Pyrosequencing™ was performed using the primer as shown in Example 1. The resulting peak heights were plotted versus expected allele frequency, and a linear relationship between the 2 was demonstrated. The experimental conditions are as set out in Example 5.

FIG. **4***b* depicts graphically relative peak heights from a Pyrosequencing reaction plotted against allele frequency. The SNP analysed was SNPE7. 5 pmol pooled DNA PCR product was incubated with 17.5 μl magnetic beads, and Pyrosequencing™ was performed using the primer as shown in Example 1. The resulting peak heights were plotted versus expected allele frequency, and a linear relationship between the 2 was demonstrated. The experimental conditions are as set out in Example 5.

FIG. **4***c* depicts graphically relative peak heights from a Pyrosequencing reaction plotted against allele frequency. The SNP analysed was SNPE4. 5 pmol pooled DNA PCR product was incubated with 17.5 μl magnetic beads, and Pyrosequencing™ was performed using the primer as shown in Example 1. The resulting peak heights were plotted versus expected allele frequency, and a linear relationship between the 2 was demonstrated. The experimental conditions are as set out in Example 5.

FIG. **5** is a further representation of FIG. **4***b*. Also depicted on this figure are the Pyrogram™ plots showing 25% C, 50% C and 75% C peaks, which are correlated to points on the linear plot. Experimental conditions are described in Example 5.

FIG. **6** depicts the obtained allele frequency results from Pyrosequencing™ for SNP 1000F and the expected allele frequency for the sample. The results are plotted as obtained allele frequency (%) versus expected allele frequencies (%). The standard line shows an imaginary pattern for an "ideal" SNP. 30 μl of PCR product was used for Pyrosequencing™, as described in Example 5.

FIG. **7** depicts the obtained allele frequency results from Pyrosequencing™ for SNP 345F and the expected allele frequency for the sample. The results are plotted as obtained allele frequency (%) versus expected allele frequencies (%). The standard line shows an imaginary pattern for an "ideal" SNP. 30 μl of PCR product was used for Pyrosequencing™, as described in Example 5. Two pools were made, with expected allele frequencies of 10% A and 26% A.

FIG. **8***a* shows DNA sequencing on pooled genomic DNA over SNP 345F (A/GGGG). 30 μl of PCR product was incubated with 10 μl magnetic beads and 20 μl of 2×BW buffer. Pyrosequencing™ was then performed on a PSQ™96 system instrument using Pyrosequencing™ SNP reagent kit. The resultant emitted light caused by nucleotide incorporation was measured and plotted as nucleotide incorporation V light emitted (RLU). For this experiment the addition of the nucleotides was such that the SNP was represented in 3 consecutive peaks (marked). The experimental conditions are as described in Example 5.

FIG. **8***b* shows DNA sequencing on pooled genomic DNA over SNP 345F (A/GGGG). 30 μl of PCR product was incubated with 10 μl magnetic beads and 20 μl of 2×BW buffer. Pyrosequencing™ was then performed on a PSQ™96 system instrument using Pyrosequencing™ SNP reagent kit. The resultant emitted light caused by nucleotide incorporation was measured and plotted as nucleotide incorporation V light emitted (RLU). For this experiment the addition of the nucleotides was such that the SNP was represented in only 2 consecutive peaks (marked). The experimental conditions are as described in Example 5.

FIG. **9** depicts the obtained mean allele frequency results from Pyrosequencing™ for SNP 471F and the expected allele frequency for the sample. The results are plotted as mean allele frequency (calculated from 10 replicates) (%) versus expected allele frequencies (%). The standard line shows an imaginary pattern for an "ideal" SNP. 30 μl of PCR product was used for Pyrosequencing™, as described in Example 5. Four pools were collated, with expected allele frequencies of 68.7%, 78.6%, 91.7% and 95.5% C.

FIG. **10***a* depicts the allele frequency obtained via Pyrosequencing™ compared to the expected allele frequency for that pool, in percentage. 3 artificial oligonucleotides were investigated, and the results for all 3 oligonucleotides are depicted. The plot is obtained allele frequency vs expected allele frequency. The oligonucleotides were used at a concentration of 1 pmol/μl, and Pyrosequencing was performed as described in Example 5. The mean frequency was calculated from 10 replicate experiments.

FIG. **10***b* depicts the results obtained for oligo 1, as shown on FIG. **10***a*.

FIG. **10***c* depicts the results obtained for oligo 2, as shown on FIG. **10***a*.

FIG. **10***d* depicts the results obtained for oligo 3, as shown on FIG. **10***a*.

FIG. **11***a* represents graphically estimated allele frequency for the C allele of SNP 465R versus template amount in the PCR reaction, the allele frequency was determined via Pyrosequencing™. 4 pools with the same allele frequency were set up using long, 10 ng, 1 ng, 0.1 ng and 0.05 ng of genomic DNA prior to PCR. The experimental conditions are as described in Example 6. The expected frequency of the C allele for each of the 4 pools was 31%.

FIG. **11***b* represents graphically estimated allele frequency for the C allele of SNP 465R versus template amount in the PCR reaction, the allele frequency was determined via Pyrosequencing™. 4 pools with the same allele frequency were set up using long, 10 ng, 1 ng, 0.1 ng and 0.05 ng of genomic DNA prior to PCR. The experimental conditions are as described in Example 6. The expected frequency of the C allele for each of the 4 pools was 12.5%.

FIG. **11***c* represents graphically estimated allele frequency for the C allele of SNP 465R versus template amount in the PCR reaction, the allele frequency was determined via Pyrosequencing™. 4 pools with the same allele frequency were set up using 10 ng, 1 ng, 0.1 ng and 0.05 ng of genomic DNA prior to PCR. The experimental conditions are as described in Example 6. The expected frequency of the C allele for each of the 4 pools was 19%.

FIG. **11***d* represents graphically estimated allele frequency for the C allele of SNP 465R versus template amount in the PCR reaction, the allele frequency was determined via Pyrosequencing™. 4 pools with the same allele frequency were set up using long, 10 ng, 1 ng, 0.1 ng and 0.05 ng of genomic DNA prior to PCR. The experimental conditions are as described in Example 6. The expected frequency of the C allele for each of the 4 pools was 6%.

FIG. **12** represents graphically estimated allele frequency obtained via Pyrosequencing™ versus peak height obtained

via Pyrosequencing™. 4 different SNPs were investigated—481R, 486R, 460R and 470R. The expected allele frequencies were as follows: 470R—55% A, 481R—19.5% G, 486R—12.5% C and 460R, 6% G. Pyrosequencing™ was performed on 5 different amounts of PCR product of pooled DNA: 30 μl, 20 μl, 15 μl, 10 μl and 5 μl. The experimental conditions are as described in Example 6.

DETAILED DESCRIPTION OF THE INVENTION

Accordingly, the present invention provides a method of determining the frequency of an allele in a population of nucleic acid molecules, said method comprising:

pooling the nucleic acid molecules of said population, performing primer extension reactions using a primer which binds at a predetermined site located in said nucleic acid molecules, and obtaining a pattern of nucleotide incorporation.

Further, the present invention provides a method of determining the amount of an allele in a sample of nucleic acid molecules, said method comprising:

performing primer extension reactions on said nucleic acid molecules, using a primer which binds at a predetermined site located in at least one said molecule, and determining which and/or how many nucleotides are incorporated in said reaction, and analysing said nucleotide incorporation information thus obtained in order to determine the amount of occurrence of said allele in said sample.

The nucleic acid molecules mentioned in the allele quantification method above may be obtained from one individual, i.e. an individual who is suspected to have additional genes, chromosomes or genomes present (i.e. trisomy or duplication of chromosomes). The nucleic acid molecules of the sample thus contain, or are suspected to contain, 3 or more alleles (e.g. 3, 4, 5 alleles). The method of the invention thus quantifies the number of alleles present (and hence the number of nucleic acid molecules which contain them), thus allowing diagnosis of gene, chromosome or whole genome duplications (or other multiplications). Thus, for example, an individual with a particular trisomy will contain 3 copies of chromosomes instead of 2. Accordingly a sample from that individual will contain 3 nucleic acid molecules corresponding to, or deriving from that chromosome, rather than two. By quantifying the amount of an allele present in that molecule, the amount of the molecule, and hence the chromosome number may be determined. In analogous fashion other duplications (i.e. replications or multiplications or indeed loss of chromosomes (e.g. chromosome number abnormalities), genes, genomes or other nucleotide sequences) may be determined. In this method an allelic variant or a particular allele may be used as a marker of a particular gene or chromosome or gene or other genetic (i.e. nucleotide) sequence it is desired to quantify.

Primer extension reactions are thus performed using the nucleic acid molecules in the pool or sample as templates. The primer, which is designed or selected to bind at a particular site in the template (e.g. adjacent, or upstream or downstream of, e.g. near to a test SNP of interest) is simply added to the sample (e.g. pooled sample for allele frequency determination) and will bind to the different template molecules present. Primer extension reactions (e.g. performed using polymerase and added nucleotides) are thus performed simultaneously or substantially simultaneously. By detecting the incorporation or non-incorporation of a given added nucleotide, a "pattern" of nucleotide incorporation may be determined which may be used to provide data which is informative on the nature of the alleles in question, and on their frequency, or occurrence (e.g. presence or absence) in the tested population. Thus, data, which may be quantitative and/or qualitative, may be obtained which may be correlated to (or which may provide information relating to) the frequency of an SNP allele (i.e. the "test" or "target" SNP or "test" or "target" allele) in the tested population. In other words, the method of the invention may be used to obtain quantitative and/or qualitative data on nucleotide incorporation relating to the SNP or allelic variant of interest.

As will be described further below, the nucleotide incorporation may be detected in various ways, and different ways of performing the primer extension reaction are possible. For example, the different nucleotides (i.e. having the different bases (e.g. A, T, C or G) may be added together, in a form in which they are distinguishable from one another (e.g. by being provided with distinguishable detectable moieties e.g. labels). More preferably however, different nucleotides may be added individually, e.g. in turn (i.e. sequentially) and the incorporation or non-incorporation of each nucleotide determined. As will be described further below, depending on the detection system selected, and/or on the target allele/SNP under test, it may not be necessary to add or use all four nucleotides (i.e. all of A, T, C or G), but a desired selection thereof.

The term "allele frequency" as used herein refers to the level or occurrence, or more particularly, the percentage of a particular allele in a given population. An allele is one of several alternative forms of a gene or nucleotide sequence at a specific chromosomal location. An allele can be any genetic variation at a given position within the nucleic acid sample. As explained above, an allele may be represented by one or more base changes at a given locus (e.g. an SNP). At each autosomal locus a diploid individual possesses 2 alleles, one maternally inherited, the other paternally. Particularly, the allele frequency determination method of the invention includes methods for determining SNP or other allelic variant allele frequencies. Each diploid individual possesses 2 alleles at a given locus. If both of the alleles are identical, the individual is homozygous for that locus. If the alleles are different, the individual is heterozygous for that locus. In the method of the invention, the frequency of each allele in the population is determined, but data on the genotype (i.e. whether the individual is homozygous for a particular allele) of a particular individual in the population will not be determined by this method.

Where allele frequency determination (i.e. allele quantification) is performed on a single sample (e.g. a sample from a single individual, for example with suspected chromosome number abnormality (e.g. trisomy) no pooling is needed.

The term "biallelic marker" as used herein refers to a genetic marker which only occurs in two forms in the population. SNPs are normally biallelic markers, although some triallelic or tetra-allelic SNPs are known and therefore the method of the invention will determine the frequency of each of the two or three or four possible alleles ("allelic variants") in a given population.

The term "population" as used herein refers to a collection of individuals, or a group. For example, the individual could be a cell, in which case the population would be a collection of cells from one or more entities or from different sites of a multi-cellular organism, or indeed cells at different stages (e.g. life stages of an organism or at different stages of the cell cycle) or a population of cells of a unicellular organism (e.g. a prokaryote). Alternatively, the individual may be a cell component, i.e. mitochondria. Further, the population may comprise individuals of the same species (i.e. humans,

domestic animals, livestock animals, plants etc.) who may or may not inhabit the same areas, region or country. The population may be selected on the basis of nationality, ethnic background, disease status, or on the basis of any other classification. Further, the population may be selected on the basis of disease susceptibility (i.e. at risk of developing cardiovascular disease) or on the basis of lack of susceptibility to disease. Familial populations (i.e. all living members of one family group or sub-division of a family, e.g. particular sibling groups) may be used. A "population" may also comprise a sample of a particular cell type or tissue from different individuals e.g. a tumour, or particular organ etc. Thus, a population may comprise nucleic acid molecules derived from a particular tissue type or diseased tissue from a number of different individuals having or exhibiting that tissue or cell type, or tumour etc. The "population" as defined herein may comprise any number of individuals, from 2 or more, to several thousand (i.e. 2 to 10,000, 2 to 8,000, 2 to 5,000).

For the analysis of gene, chromosome or genome number (i.e. quantification or multiplication detection), the individual is defined as "the population". The sample from an individual may contain a variant amount or number of a given (e.g. target) nucleic acid molecule. This allele quantification can be performed on single samples which may contain a variable number or amount of a target nucleic acid molecule (target allele).

The term "pooled nucleic acid molecules" as used herein refers to the pooling of nucleic acid molecules into one reaction mixture from all individuals of a given population (i.e. the adding together of the different or individual nucleic acid samples to create a pooled sample). Therefore, multiple individual nucleic acid molecules are pooled prior to genetic analysis. Pooling of nucleic acid molecules is sample size independent, i.e. independent of the number of samples comprising the pool.

"Multiple" as used herein means two or more e.g. 3, 4, 5, 6, 8, 10 or more, or 100, 200, 500, 1000, 2000, 5000 or 10000 or more.

Conveniently, the nucleic acid molecule may be DNA, although determining the allele frequency of RNA (e.g. mRNA) is also within the invention. If it is desired to use a RNA sample, the method may additionally include the step of generating cDNA from the RNA template, conveniently by using reverse transcriptase. Alternatively, if desired, the primer extension reactions may be performed directly on RNA templates.

The target nucleic acid may thus be any nucleic acid, isolated or synthetic, in any desired or convenient form. It may thus be genomic DNA, or isolated mRNA which may be used directly for analysis by the method of the invention, or it may be a nucleic acid product derived therefrom (or corresponding thereto), e.g. by synthesis, such as cDNA as mentioned above, or an amplification product (e.g. PCR amplicon), clones or library products etc.

In carrying out the method of the invention, a primer specific for the allele of interest is provided which binds to the nucleic acid molecules at a predetermined site. The primer is designed or selected so that when the primer extension reaction is performed, the primer is extended over the allele (or SNP) in the nucleic acid. In other words, the primer binds to the nucleic acid molecule at, or near to (e.g. within 1 to 20, 1 to 10 or 1 to 6 bases), the allele/SNP.

It will be understood that in order to perform the invention the primer binding site should be available in all individual nucleic acid molecules in the pooled population. Such primer binding sites will therefore advantageously lie in regions which are common to, or substantially conserved between the different individuals in the population. This may readily be achieved by selecting the primer binding site to lie in conserved/semi-conserved regions as discussed above.

It will therefore be understood that in the pooled nucleic acid, there will generally be 2 "allelic variants" present for each SNP. Thus, at a given polymorphic position, the nucleotide may be either one or two possible bases. In the case of triallelic SNP, there will be one of 3 possible bases. In the case of tetra-allelic SNPs there will be one or two of four possible bases.

Preferably, the polymorphic position is not sequenced within a homopolymeric stretch in either allelic variant. As used herein a homopolymeric stretch is defined as a stretch of nucleic acid which contains two or more (i.e. 3 or more, 4 or more or 5 or more) consecutive identical nucleotides (i.e. $GC_{AAA}T$). However, primers can be designed to avoid sequencing the homopolymeric stretch whilst obtaining data on the allele frequency. Therefore, with well designed primers, estimating allele frequencies of alleles present in homopolymeric stretches is within the scope of the invention. It is possible to design the primer in order to avoid sequencing the repeated bases. The extension primer can thus be designed to cover the homopolymeric region.

Further, by the use of appropriate controls or conditions, and depending on the detection method chosen, it is possible to determine the frequency of an allele if the SNP is in a homopolymeric stretch.

The primer extension reactions conveniently may be performed by sequentially adding nucleotides to the reaction mixture (i.e. polymerase and primer/template mixture). Advantageously, the different nucleotides are added in known predetermined order. As each nucleotide is added, it may be determined whether or not nucleotide incorporation takes place.

Advantageously, as described in more detail below, the amount of nucleotide incorporated (i.e. how many nucleotide residues) may be determined. Such a quantitative embodiment, wherein nucleotide incorporation is determined quantitatively, represents a preferred aspect of the invention.

In this manner, sequencing data may be obtained for the polymorphic position in all nucleic acid molecules in the pooled samples. This sequencing data comprises the base identity (i.e. sequence) of the particular SNP residue, together with quantitative data on how many nucleotides of each type have been incorporated. In other words, the data corresponds to the allele frequency for the given SNP. The allele frequency may thus readily be calculated using the quantitative values obtained for nucleotide incorporation during primer extension wherein the primer is extended over the polymorphic position.

Thus, by identifying how much of each nucleotide is incorporated at the polymorphic site in a primer extension reaction, it is possible to calculate the frequency of each allele.

In order to perform the invention, it may be advantageous or convenient first to amplify the nucleic acid molecule by any suitable amplification method known in the art. The target nucleic acid would then be an amplicon. Suitable in vitro amplification techniques include any process which amplifies the nucleic acid present in the reaction under the direction of appropriate primers. The amplicon method may thus preferably be PCR, or any of the various modifications thereof e.g. the use of nested primers, although it is not limited to this method. Those skilled in the art will appreciate that other amplification procedures may also be used,

such as Self-sustained Sequence Replication (3SR), NASBA, the Q-beta replicase amplification system and Ligase chain reaction (LCR) (see for example Abramson and Myers (1993) Current Opinion in Biotech., 4: 41–47). If PCR is used to amplify the nucleic acid, suitable primers, are designed to ensure that the region of interest within the nucleic acid sequence (i.e. the region containing the SNP), is amplified. PCR can also be used for indiscriminate amplification of all nucleic acid sequences, allowing amplification of essentially all sequences within the sample for study (i.e. total nucleic acid). Linker-primer PCR is particularly suitable for indiscriminate amplification, and uses double stranded oligonucleotide linkers with a suitable overhanging end, which are ligated to the ends of target nucleic acid fragments. Amplification is then conducted using oligonucleotide primers which are specific for the linker sequences. Alternatively, completely random oligonucleotide primers may be used in conjunction with DOP-PCR (degenerate oligonucleotide-primed) to amplify all the nucleic acid within a sample.

One or more of the amplification primers used in the amplification reaction, may be subsequently used as an "extension primer", but this will preferably be a different primer. It will be appreciated that the sequence and length of the oligonucleotide amplification and extension primers to be used in the amplification and extension steps, respectively, will depend on the sequence of the target nucleic acid, the desired length of amplification or extension product, the further functions of the primer (i.e. for immobilization) and the method used for amplification and/or extension. Appropriate primers may readily be designed applying principles and techniques well known in the art.

Advantageously, as mentioned above, an extension primer will bind substantially adjacent (e.g. within 1–20, 1–10 or 1–6, preferably within 1–3 bases), or exactly adjacent to the SNP of the target nucleic acid molecules and may be complementary to a conserved or semi-conserved region of the nucleic acid molecules. In order for the method of the invention to be performed, knowledge of the sequence surrounding the allele (e.g. of the conserved or semi-conserved region) is required in order to design an appropriate complementary extension primer. The specificity is achieved by virtue of complementary base pairing. For all embodiments of the invention, primer design may be based upon principles well known in the art. It is not necessary for the extension or amplification primer to have absolute complementarity to the binding site, but this is preferred to improve the specificity of binding.

The extension primer may be designed to bind to the sense or anti-sense strand of the target nucleic acid.

The "primer extension" reaction according to the invention includes all forms of template-directed polymerase-catalysed nucleic acid synthesis reactions. Conditions and reagents for primer extension reactions are well known in the art, and any of the standard methods, reagents and enzymes etc. may be used in this step (see e.g. Sambrook et al., (eds), Molecular Cloning: a laboratory manual (1989), Cold Spring Harbor Laboratory Press). Thus, the primer extension reaction at its most basic, is carried out in the presence of primer, deoxynucleotides (dNTPs) and a suitable polymerase enzyme e.g. T7 polymerase, Klenow or Sequenase Ver 2.0 (USB USA), or indeed any suitable available polymerase enzyme. As mentioned above, for an RNA template, reverse transcriptase may be used. Conditions may be selected according to choice, having regard to procedures well known in the art.

The primer is thus subjected to a primer-extension reaction in the presence of a nucleotide, whereby the nucleotide is only incorporated if it is complementary to the base immediately adjacent (3') to the primer position. The nucleotide may be any nucleotide capable of incorporation by a polymerase enzyme into a nucleic acid chain or molecule. Thus, for example, the nucleotide may be a deoxynucleotide (dNTP, deoxynucleoside triphosphate) or dideoxynucleotide (ddNTP, dideoxynucleoside triphosphate). Thus, the following nucleotides may be used in the primer-extension reaction: guanine (G), cytosine (C), thymine (T) or adenine (A) deoxy- or dideoxy-nucleotides. Therefore, the nucleotide may be dGTP (deoxyguanosine triphosphate), dCTP (deoxycytidine triphosphate), dTTP (deoxythymidine triphosphate) or dATP (deoxyadenosine triphosphate). As discussed further below, suitable analogues of DATP, and also for dCTP, dGTP and dTTP may also be used. Thus, modified nucleotides, or nucleotide derivatives (e.g. chemically modified nucleotides) may be used so long as they are capable of incorporation by a polymerase enzyme. Dideoxynucleotides may also be used in the primer-extension reaction. The term "dideoxynucleotide" as used herein includes all 2'-deoxynucleotides in which the 3' hydroxyl group is modified or absent. Dideoxynucleotides are capable of incorporation into the primer in the presence of the polymerase, but cannot enter into a subsequent polymerisation reaction, and thus function as a "chain terminator". It will therefore be appreciated that in embodiments of the invention which rely on sequential nucleotide addition the use of chain terminating nucleotides is to be avoided (although so-called "false" or "labile" terminators might be used in which the 3' blocking group may be removed following incorporation. Such modified nucleotides are known and described in the art). However, in some embodiments of the invention it may be advantageous to use chain terminating nucleotides whereby it is desired to terminate sequencing of one allele after incorporation of the chain terminating nucleotide, but more sequence information is required for the other allele.

If the nucleotide is complementary to the target base, the primer is extended by one nucleotide, and inorganic pyrophosphate is released. As discussed further below, in a preferred method, the inorganic pyrophosphate may be detected in order to detect the incorporation of the added nucleotide. For the SNP of interest, the addition of two nucleotides will be sufficient to generate allele frequency information. The primer bound to one allelic variant will be extended by 1 nucleotide upon addition of the nucleotide which base pairs to the nucleotide in the polymorphic position. The primer bound to the other allelic variant will therefore not be extended by addition of this nucleotide. This primer will be extended in the next round of nucleotide addition, which should be designed to be a complementary base to the allelic variant (i.e. if the allelic variant is C, a G should be added). Different nucleotides may be added sequentially, advantageously in known order, as discussed above, to reveal the nucleotides which are incorporated for each extension primer. Accordingly, determining the number of nucleotides incorporated for each nucleotide addition, will reveal the frequency of that allele corresponding to nucleotide incorporation and hence contribute to the calculation of allele frequency.

Hence, a primer extension protocol may involve annealing a primer as described above, adding a nucleotide, performing a polymerase-catalysed primer extension reaction, detecting the presence or absence of incorporation of said nucleotide (and advantageously also determining the amount of each nucleotide incorporated) and repeating the

nucleotide addition and primer extension steps etc. one or more times. As discussed above, single (i.e. individual) nucleotides may be added successively to the same primer-template mixture.

In order to permit the repeated or successive (iterative) addition of nucleotides in a primer-extension procedure, the previously-added nucleotide must be removed. This may be achieved by washing, or more conveniently, by using a nucleotide-degrading enzyme, for example as described in detail in WO98/28440.

Accordingly, in a principal embodiment of the present invention, a nucleotide degrading enzyme is used to degrade any unincorporated or excess nucleotide. Thus, if a nucleotide is added which is not incorporated (because it is not complementary to the target base), or any added nucleotide remains after an incorporation event (i.e. excess nucleotides) then such unincorporated nucleotides may readily be removed by using a nucleotide-degrading enzyme. This is described in detail in WO98/28440.

The term "nucleotide degrading enzyme" as used herein includes any enzyme capable of specifically or non-specifically degrading nucleotides, including at least nucleoside triphosphates (NTPs), but optionally also di- and monophosphates, and any mixture or combination of such enzymes, provided that a nucleoside triphosphatase or other NTP-degrading activity is present. Where a chain terminating nucleotide is used (e.g. a dideoxy nucleotide is used), the nucleotide degrading enzyme should also degrade such a nucleotide. Although nucleotide-degrading enzymes having a phosphatase activity may conveniently be used according to the invention, any enzyme having any nucleotide or nucleoside degrading activity may be used, e.g. enzymes which cleave nucleotides at positions other than at the phosphate group, for example at the base or sugar residues. Thus, a nucleoside triphosphate degrading enzyme is essential for the invention. Nucleoside di- and/or mono-phosphate degrading enzymes are optional and may be used in combination with a nucleoside tri-phosphate degrading enzyme.

The preferred nucleotide degrading enzyme is apyrase, which is both a nucleoside diphosphatase and triphosphatase, catalysing the reactions NTP→NDP+Pi and NDP→NMP+Pi (where NTP is a nucleoside triphosphate, NDP is a nucleoside diphosphate, NMP is a nucleotide monophosphate and Pi is inorganic phosphate). Apyrase may be obtained from the Sigma Chemical Company. Other possible nucleotide degrading enzymes include Pig Pancreas nucleoside triphosphate diphosphorydrolase (Le Bel et al., 1980, J. Biol. Chem., 255, 1227–1233). Further enzymes are described in the literature.

The nucleotide-degrading enzyme may conveniently be included during the polymerase (i.e. primer extension) reaction step. Thus, for example the polymerase reaction may conveniently be performed in the presence of a nucleotide-degrading enzyme. Although less preferred, such an enzyme may also be added after nucleotide incorporation (or non-incorporation) has taken place, i.e. after the polymerase reaction step.

Thus, the nucleotide-degrading enzyme (e.g. apyrase) may be added to the polymerase reaction mixture (i.e. target nucleic acid, primer and polymerase) in any convenient way, for example prior to or simultaneously with initiation of the reaction, or after the polymerase reaction has taken place, e.g. prior to adding nucleotides to the sample/primer/polymerase to initiate the reaction, or after the polymerase and nucleotide are added to the sample/primer mixture.

Conveniently, the nucleotide-degrading enzyme may simply be included in the reaction mixture for the polymerase reaction, which may be initiated by the addition of the nucleotide.

According to the present invention, detection of nucleotide incorporation can be performed in a number of ways, such as by incorporation of labelled nucleotides which may subsequently be detected, or by using labelled probes which are able to bind to the extended sequence.

The method may be performed using a Sanger sequencing method combined with a standard detection strategy, e.g. electrophoresis or mass spectrometry to analyse, or determine, nucleotide incorporation. However, it is preferred to use a sequencing-by-synthesis method, due to the fact that the extension reactions are quantitative, i.e. that the nucleotide incorporation may be determined quantitatively. As mentioned above, sequencing-by-synthesis methods are disclosed extensively in U.S. Pat. No. 4,863,849, which discloses a number of ways in which nucleotide incorporation may be determined or detected, e.g. spectrophotometrically or by fluorescent detection techniques, for example by determining the amount of nucleotide remaining in the added nucleotide feedstock, following the nucleotide incorporation step. Alternatively, labelled nucleotides may be utilised in the nucleotide incorporation step. Such labelled nucleotides may be chain terminating or capable of further extension. The nucleotide incorporated may be identified and the label removed or neutralised prior to the incorporation of the next labelled nucleotide. Such a method is described in U.S. Pat. No. 6,087,095 of Rosenthal et al. This patent also describes sequencing-by-synthesis on a solid phase (e.g. beads). The label may be a fluorescent label or a radioactive label.

The preferred method of sequencing-by-synthesis is however a pyrophosphate detection-based method.

Preferably, therefore, nucleotide incorporation is detected by detecting PPi release, preferably by luminometric detection, and especially by bioluminometric detection.

PPi can be determined by many different methods and a number of enzymatic methods have been described in the literature (Reeves et al., (1969), Anal. Biochem., 28, 282–287; Guillory et al., (1971), Anal. Biochem., 39, 170–180; Johnson et al., (1968), Anal. Biochem., 15, 273; Cook et al., (1978), Anal. Biochem. 91, 557–565; and Drake et al., (1979), Anal. Biochem. 94, 117–120).

It is preferred to use luciferase and luciferin in combination to identify the release of pyrophosphate since the amount of light generated is substantially proportional to the amount of pyrophosphate released which, in turn, is directly proportional to the amount of nucleotide incorporated. The amount of light can readily be estimated by a suitable light sensitive device such as a luminometer. Thus, luminometric methods offer the advantage of being able to be quantitative.

Luciferin-luciferase reactions to detect the release of PPi are well known in the art. In particular, a method for continuous monitoring of PPi release based on the enzymes ATP sulphurylase and luciferase has been developed (Nyrén and Lundin, Anal. Biochem., 151, 504–509, 1985; Nyrén P., Enzymatic method for continuous monitoring of DNA polymerase activity (1987) Anal. Biochem Vol 167 (235–238)) and termed ELIDA (Enzymatic Luminometric Inorganic Pyrophosphate Detection Assay). The use of the ELIDA method to detect PPi is preferred according to the present invention. The method may however be modified, for example by the use of a more thermostable luciferase (Kaliyama et al., 1994, Biosci. Biotech. Biochem., 58, 1170–1171) and/or ATP sulfurylase (Onda et al., 1996,

Bioscience, Biotechnology and Biochemistry, 60:10, 1740–42). This method is based on the following reactions:

$$PPi + APS \xrightarrow{\text{ATP sulphurylase}} ATP + SO_4^{2-}$$

$$ATP + \text{luciferin} + O_2 \xrightarrow{\text{luciferase}} AMP + PPi + \text{oxyluciferin} + CO_2 + hv$$

(APS=adenosine 5'-phosphosulphate)

Reference may also be made to WO 98/13523 and WO 98/28448, which are directed to pyrophosphate detection-based sequencing procedures, and disclose PPi detection methods which may be of use in the present invention.

In a PPi detection reaction based on the enzymes ATP sulphurylase and luciferase, the signal (corresponding to PPi released) is seen as light. The generation of the light can be observed as a curve known as a Pyrogram™. Light is generated by luciferase action on the product, ATP (produced by a reaction between PPi and APS (see below) mediated by ATP sulphurylase) and, where a nucleotide-degrading enzyme such as apyrase is used, this light generation is then "turned off" by the action of the nucleotide-degrading enzyme, degrading the ATP which is the substrate for luciferase. The slope of the ascending curve may be seen as indicative of the activities of DNA polymerase (PPi release) and ATP sulphurylase (generating ATP from the PPi, thereby providing a substrate for luciferase). The height of the signal is dependent on the activity of luciferase, and the slope of the descending curve is, as explained above, indicative of the activity of the nucleotide-degrading enzyme. As explained below, in a Pyrogram™ in the context of a homopolymeric region, peak height is also indicative of the number of nucleotides incorporated for a given nucleotide addition step. Then, when a nucleotide is added, the amount of PPi released will depend upon how many nucleotides (i.e. the amount) are incorporated, and this will be reflected in the peak height.

The use of pyrophosphate detection-based sequencing methods, and in particular those based on the ELIDA detection enzymes, is particularly advantageous in the present invention; the correlation between signals obtained in such methods (i.e. peak heights) and SNP allele frequencies has been shown to be excellent, and the accuracy of the results obtained surprisingly high. Frequencies as lows as 5% for one allele have been determined with reasonable accuracy in pools of samples.

Advantageously, by including the PPi detection enzyme(s) (i.e. the enzyme or enzymes necessary to achieve PPi detection according to the enzymatic detection system selected, which in the case of ELIDA, will be ATP sulphurylase and luciferase) in the polymerase reaction step, the method of the invention may readily be adapted to permit extension reactions to be continuously monitored in real-time, with a signal being generated and detected, as each nucleotide is incorporated.

Thus, the PPi detection enzymes (along with any enzyme substrates or other reagents necessary for the PPi detection reaction) may simply be included in the polymerase reaction mixture.

A potential problem which has previously been observed with PPi-based sequencing methods is that dATP, used in the chain extension reaction, interferes in the subsequent luciferase-based detection reaction by acting as a substrate for the luciferase enzyme. This may be reduced or avoided by using, in place of deoxyadenosine triphosphate (ATP), a DATP analogue which is capable of acting as a substrate for a polymerase but incapable of acting as a substrate for a PPi-detection enzyme. Such a modification is described in detail in WO98/13523.

The term "incapable of acting" includes also analogues which are poor substrates for the detection enzymes, or which are substantially incapable of acting as substrates, such that there is substantially no, negligible, or no significant interference in the PPi detection reaction.

Thus, a further preferred feature of the invention is the use of a dATP analogue which does not interfere in the enzymatic PPi detection reaction but which nonetheless may be normally incorporated into a growing DNA chain by a polymerase. By "normally incorporated" is meant that the nucleotide is incorporated with normal, proper base pairing. In the preferred embodiment of the invention where luciferase is a PPi detection enzyme, the preferred analogue for use according to the invention is the [1-thio] triphosphate (or α-thiotriphosphate) analogue of deoxy ATP, preferably deoxyadenosine [1-thio] triphosphate, or deoxyadenosine-thiotriphosphate (dATP αS) as it is also known. dATP αS, along with the α-thio analogues of dCTP, dGTP and dTTP, may be purchased from Amersham Pharmacia. Experiments have shown that substituting dATP with dATP αS allows efficient incorporation by the polymerase with a low background signal due to the absence of an interaction between dATP αS and luciferase. False signals are decreased by using a nucleotide analogue in place of dATP, because the background caused by the ability of dATP to function as a substrate for luciferase is eliminated. In particular, an efficient incorporation with the polymerase may be achieved while the background signal due to the generation of light by the luciferin-luciferase system resulting from DATP interference is substantially decreased. It has been noted by the inventors that the use of dATP αS can lead to higher peaks than the use of dATP. The peak height is consistently higher, and thus if dATP αS is used, the actual 'peak height' can be calculated via a 'peak height reduction'. The dNTP αS analogues of the other nucleotides may also be used in place of the other dNTPs.

The step of detecting nucleotide incorporation by detecting PPi release results in a signal indicative of the amount of pyrophosphate released, and hence the amount of nucleotide incorporated.

In the method of the invention, the primer-extension reaction is performed simultaneously for each nucleic acid molecule in the reaction mixture. Thus, for every nucleotide addition to the reaction mixture, multiple nucleotides may be incorporated into the extended primers. The signal generated in the pyrophosphate detection step will therefore be indicative of the number of nucleotides incorporated in the primer-extension step for the combination of all primers bound to the template nucleic acid. The size of the signal (i.e. the height of each peak) can therefore be correlated directly to the number of incorporated nucleotides. Typically, the primer needs only to be subjected to 1 to 20, preferably 1 to 10, e.g. 1 to 5 and most preferably 2 to 4 cycles of nucleotide addition.

It will be understood that the order of nucleotide addition in the reaction mixture can be tailored to each SNP to ensure that the relevant allele frequency is obtained efficiently and accurately. For example, if the 2 possible allelic nucleotides are C or T (or vice versa), the order of nucleotide addition when extending the primer over the polymorphic site may be C followed by T, using the methods as described previously.

Therefore, the peaks showing nucleotide incorporation for the allelic variant bases should preferably be adjacent to each other, facilitating calculation of the allele frequencies.

As mentioned previously, the allele variants are preferably not sequenced in a homopolymeric stretch of 3 or more identical bases. It will be clear that the peak height in such a situation will represent not only the nucleotide incorporation relating to the polymorphic position, but will also represent the incorporation of 2 or more nucleotides further downstream of the polymorphism. Thus, the number of nucleotides incorporated will also reflect the number of nucleotides present in the homopolymeric region, which will be the same for each allelic variant. Therefore, it is advisable to avoid performing allele frequency determinations on SNPs wherein one allelic variant lies within a homopolymeric stretch of three or more identical bases, unless a primer can be designed as described previously.

It will be understood that in order to obtain accurate and reliable data relating to the frequency of an allele in a population, it will be preferable to use the same amount of nucleic acid for each individual in the population in the reaction mixture. Therefore, it may be necessary to calibrate the samples prior to pooling. Thus, it forms a preferred aspect of the invention to measure or determine the concentration of the nucleic acid in the sample prior to pooling. Any standard technique may be used to effect the measurement/determination of nucleic acid concentration, such as gel electrophoresis and spectrophotometry. However, these methods are not without their drawbacks, as they rely upon having a significant sample of nucleic acid to use for concentration determination. A further aspect of this invention is thus using a primer-extension reaction to calibrate the nucleic acid concentrations prior to pooling.

In order to perform primer extension reactions to calculate the concentration of nucleic acid in a sample, it will first be necessary to select a suitable SNP. A suitable SNP for such analysis will not be present in a homopolymeric sequence and will not be preferentially amplified in any PCR-type reactions. Further, the SNP should be chosen such that it gives no background signals in a primer-extension reaction, and that the signals, e.g. peak height, (see before) are even. Preferably, each of the individuals has a known sequence (genotype) at this SNP. If not, the sequence (genotype) can be determined using standard sequence-by-synthesis reaction means. One reference sample (Ref 1) is selected as the main reference from one of the homozygotes, another reference sample (Ref 2) is selected from the other homozygote, and are pooled, and the method of the invention as previously described may be carried out. The results of the primer extension reactions enable the relative concentrations of each reference sample to be calculated, as the signals (e.g. peak heights) (see before) are directly related to the amount of nucleotide incorporation. To measure the concentration of the rest of the samples in the population, these are pooled individually with one of the reference samples. Heterozygote samples should be paired with one of the homozygote references, and then analysed as mentioned previously. Thus, as the concentration of the reference sample is known, the concentration of the sample pooled with the reference sample can be easily calculated. Homozygote samples should be pooled with the other homozygote reference sample (i.e. pair AA with CC, not AA with AA).

The peak height for allele 1 (i.e. A) and the peak height for allele 2 (i.e. C) are recorded, and the following calculations are performed (for an allele not present in a homopolymer stretch):

$$Y = \frac{\text{Peak Height (allele 1)}}{\text{Peak Height (allele 1)} + \text{Peak height (allele 2)}}$$

where Y is the frequency of allele 1. The concentration in the sample is calculated by multiplying the concentration of the reference by a concentration factor (X). Therefore, X must be calculated. X is in relation to the reference sample used. If the sample is heterozygous, X is calculated in the following way:

$$X = \frac{2Y}{1-2Y}$$

However, if the sample is homozygous, the following calculation is used:

$$X = \frac{Y}{1-Y}$$

Thus, once it has been decided what volume of one of the reference samples is to be used in the pool, the volume of samples to be added to the pool is calculated by dividing the volume for the reference with the X value for each sample i.e.

$$\text{volume (sample } n) = \frac{\text{volume (ref 1)}}{X \text{ (sample } n)}$$

Alternatively or additionally, once it has been decided what volume of one of the reference samples is to be used in the pool, the volume of the second reference sample is set by dividing the volume of reference 1 with the concentration factor (X) of reference 2.

$$\text{Volume (reference 2)} = \frac{\text{Volume (reference 1)}}{X \text{ (reference 2)}}$$

From these 2 volumes (reference 1 and reference 2) the volumes of samples to be added to the pool is calculated by dividing the volume for the reference with the X value for each sample. It is important to use the correct reference for each sample (i.e. the reference this sample has been compared to).

$$\text{Volume (sample } n) = \frac{\text{Volume (ref 1 or 2)}}{X \text{ (sample } n)}$$

Thus, although different volumes are used for each sample, the amount of nucleic acid from each individual will be the same. Calculations have been performed in Example 1.

The uniformity of nucleic acid amount of different individuals in the population (i.e. in the individual nucleic acid samples which are pooled) may vary, depending on the source and nature of the nucleic acid, and indeed the

importance of such uniformity (and hence the need for calibration) may also vary, depending on the nucleic acid samples used. Thus, when using pooled genomic DNA samples, uniformity of DNA concentration between individual samples has been found to be of more importance and it is preferred first to calibrate the sample concentration for optimum results. However, calibration is not absolutely necessary and the concentration of the nucleic acid in the sample may be estimated by standard methods.

The calibration procedure will be of particular interest, if it is important to know the exact allele frequencies in a pool, or if the pool consists of a few samples and/or there are large differences in the individual DNA concentrations.

The amount of template nucleic acid from the pool of nucleic acid used for amplification has been found by the inventors under certain circumstances to be important when performing allele frequency studies. In order to obtain reproducible results, at least 10 ng, preferably 10 to 100 ng, more preferably 10 to 50 ng and even more preferably 10 to 20 ng of nucleic acid is generally preferred. Such amounts are particularly recommended for genomic DNA but is equally applicable to cases wherein PCR products are pooled.

Generally speaking the absolute level of signal detected (e.g. peak height in a Pyrogram™), does not significantly affect the accuracy of allele frequency determinations as long as the analysed signals (e.g. peaks) are well above (i.e. distinguishably above) noise level. Generally speaking however, the lowest peak in a Pyrogram™ is ideally at least 2 RLU (relative light units) to distinguish from noise/background. Single peak heights of at least 10 or 15 RLU have generally been found to be reliable, particularly if one of the alleles is represented at a low frequency.

Preferably, the concentration of the nucleic acid in the sample is determined by a primer-extension reaction (as described previously).

Preferably, the genomic nucleic acid from all individuals in the population are pooled, and amplified prior to analysis. Suitable amplification techniques have been discussed previously. As mentioned before, the nucleic acid may be of any suitable nature. In order to increase the accuracy of allele frequency calculations, it is advisable to separate the nucleic acid pool prior to amplification into "sub-pools" (or several PCR replicates) to enable multiple allele-frequency assays of the invention to be performed for the same allele. Preferably, there are 1 or more sub-pools (i.e. 2, 3, 4, 5, 6, 7, 8, 9, 10 or more), and therefore the same study is replicated 1 or more times. As mentioned previously, there is preferably at least 10 ng of nucleic acid present in the pool prior to amplification. Calculating an average allele frequency from the sub-pools improves the accuracy of allele frequency determination when dealing with genomic or amplified nucleic acid material. The use of amplified nucleic acid in the method of the invention is also envisaged. However, less replicate allele frequency experiments need to be performed than if genomic nucleic acid is pooled.

In order for the primer-extension reaction (either for calibration or allele frequency determination) to be performed, the nucleic acid molecule, regardless of whether or not it has been amplified, is conveniently provided in a single-stranded format. The nucleic acid may be subjected to strand separation by any suitable technique known in the art (e.g. Sambrook et al., supra), for example by heating the nucleic acid, or by heating in the presence of a chemical denaturant such as formamide, urea or formaldehyde, or by use of alkali.

However, this is not absolutely necessary and a double-stranded nucleic acid molecule may be used as template, e.g. with a suitable polymerase having strand displacement activity.

Where a preliminary amplification step is used, regardless of how the nucleic acid has been amplified, all components of the amplification reaction need to be removed, to obtain pure nucleic acid, prior to carrying out the typing assay of the invention. For example, unincorporated nucleotides, PCR primers, and salt from a PCR reaction need to be removed. Methods for purifying nucleic aids are well known in the art (Sambrook et al., supra), however a preferred method is to immobilize the nucleic acid molecule, removing the impurities via washing and/or sedimentation techniques.

Optionally, therefore, the target nucleic acid may be provided with a means for immobilization, which may be introduced during amplification, either through the nucleotide bases or the primer/s used to produce the amplified nucleic acid.

To facilitate immobilization, the amplification primers used according to the invention may carry a means for immobilization either directly or indirectly. Thus, for example the primers may carry sequences which are complementary to sequences which can be attached directly or indirectly to an immobilizing support or may carry a moiety suitable for direct or indirect attachment to an immobilizing support through a binding partner.

Numerous suitable supports for immobilization of DNA and methods of attaching nucleotides to them, are well known in the art and widely described in the literature. Thus for example, supports in the form of microtitre plate (MTP) wells, tubes, dipsticks, particles, beads, fibres or capillaries may be used, made for example of agarose, sepharose, cellulose, alginate, cellulose alginate, teflon, latex or polystyrene. Advantageously, the support may comprise beads, e.g. sepharose beads produced by Amersham Biosciences (Uppsala, Sweden), or magnetic particles eg. the superparamagnetic beads produced by Dynal AS (Oslo, Norway) and sold under the trademark DYNABEADS®. Chips may be used as solid supports to provide miniature experimental systems as described for example in Nilsson et al. (Anal. Biochem. (1995), 224:400–408).

The solid support may carry functional groups such as hydroxyl, carboxyl, aldehyde or amino groups for the attachment of the primer or capture oligonucleotide. These may in general be provided by treating the support to provide a surface coating of a polymer carrying one of such functional groups, eg. polyurethane together with a polyglycol to provide hydroxyl groups, or a cellulose derivative to provide hydroxyl groups, a polymer or copolymer of acrylic acid or methacrylic acid to provide carboxyl groups or an amino alkylated polymer to provide amino groups. U.S. Pat. No. 4,654,267 describes the introduction of many such surface coatings. Alternatively, the support may carry other moieties for attachment, such as avidin or streptavidin (binding to biotin on the nucleotide sequence), DNA binding proteins (eg. the lac I repressor protein binding to a lac operator sequence which may be present in the primer or oligonucleotide), or antibodies or antibody fragments (binding to haptens eg. digoxigenin on the nucleotide sequence). The streptavidin/biotin binding system is very commonly used in molecular biology, due to the relative ease with which biotin can be incorporated within nucleotide sequences, and indeed the commercial availability of biotin-labelled nucleotides. This represents one preferred method for immobilisation of target nucleic acid molecules according to the present inven-

tion. Streptavidin-coated DYNABEADS® are commercially available from Dynal AS, and streptavidin-coated Sepharose beads are commercially available from Amersham Biosciences.

As mentioned above, immobilization may conveniently take place after amplification. To facilitate post amplification immobilisation, one or both of the amplification primers are provided with means for immobilization. Such means may comprise as discussed above, one of a pair of binding partners, which binds to the corresponding binding partner carried on the support. Suitable means for immobilization thus include biotin, haptens, or DNA sequences (such as the lac operator) binding to DNA binding proteins.

When immobilization of the amplification products is not performed, the products of the amplification reaction may simply be separated by for example, taking them up in a formamide solution (denaturing solution) and separating the products, for example by electrophoresis or by analysis using chip technology. Immobilization provides a ready and simple way to generate a single-stranded template for the extension reaction. As an alternative to immobilization, other methods may be used, for example asymmetric PCR, exonuclease protocols or quick denaturation/annealing protocols on double stranded templates may be used to generate single stranded DNA. Such techniques are well known in the art.

The method of the invention allows the determination of the frequency of an allele in a population (i.e. a group of individuals exhibiting disease or trait, a familial group, an ethnic group, a geographical group), wherein the allele assessed is a single nucleotide polymorphism (SNP) or any other allelic variant.

The method of the present invention is particularly advantageous in determining whether a particular allelic variant is linked to disease or trait. To enable such determination, 2 or more (i.e. 3 OR 4, 5, 6, 7, 8, 9 OR 10) pools of nucleic acid molecules are analyzed. One pool comes from a population exhibiting said disease or trait, whilst the second pool is selected from a population which do not exhibit said disease or trait. If the frequency of one allelic variant is greater in the 'diseased', population, this points towards the allele being associated with the disease or trait. However, it will be appreciated that the method of the invention can be performed on 1 pool in isolation.

The method of the present invention may be used to confirm whether an allelic variation is present in a population. For example, an SNP may be identified in silico (by searching databases and homologues) or identified in one population (i.e. an isolated geographical group or ethnic group), and it may be desirable to ascertain the frequency of an allele in another population (i.e. a different ethnic group or different familial group).

The method of the present invention is particularly advantageous in studies of mutations associated with cancer. In this case, the population is a sample of cells removed from a patient (i.e human, livestock animal, domestic animal or laboratory animal). In the population of cells, there will be a mixture of healthy and diseased cells, and the nucleic acid from all cells in the population will be pooled. The population can then be scanned for SNPs which are associated with diseased state in the patient, giving patient-specific information on the disease-associated allele, and the frequency of that allele in a population of cells. This type of information could be invaluable in the treatment of cancer, by aiding diagnosis and prognosis. Further, knowledge of the allele involved can allow the tailoring of treatment for the allele involved; this technology is known as pharmacogenomics. Repeated testing of a population of cells from an individual can give an estimation of the proportion of cells that are carrying the disease-associated allele. By using the method of the invention, it is possible to separate the mixed genotypes present in the mixed cell populations. This is a great advantage over prior methods where mixed genotypes were indicated due to a mixture of cell types being present. It will be understood that this technology could also be used to analyse multiploid genomes (e.g. plants). A further application of determining allele frequency from a population of cells is that loss of heterozygosity can be examined. This will detect whether a segment of chromosome has been lost in tumour tissue.

A further application of the method of the invention is testing for 'genetic drift'. Using the method of the invention, it will be possible to obtain data on a particular allele frequency within a given population at given time intervals, and determine whether over time, the frequency of an allele changes. This type of analysis will therefore involve taking nucleic acid samples from multiple generations in a population. It is thought that genetic drift is a useful indicator of evolutionary change, and the method of the invention will be able to measure such allele frequency change quickly and simply.

A further application of the method of the invention is for quantification of a gene/allele in human samples for trisomy tests (or other chromosome abnormalities or gene multiplication etc). This is important in different syndromes where one chromosome occurs in three copies instead of two as normal. A well-known syndrome is Downs Syndrome or trisomy-21. Other trisomies are trisomy-13, and 18. Other syndromes related to duplications of sex chromosomes (or other chromosome number abnormality) can also be analysed using the method of the invention. This can be performed by quantifying the number of alleles of any gene (or indeed any particular selected nucleotide sequence containing allelic variation or polymorphism) on the selected chromosome.

The method of the invention is advantageous in that it determines the exact sequence of the SNP or allelic variant, together with a direct measurement of the amount of nucleotide incorporated. The primer extension reaction generates a "pattern" indicative of nucleotide incorporation, correlated to the nucleotide added to the reaction mixture. The pattern is a cumulative picture of nucleotide incorporation for the primers bound to all of the nucleic acid molecules present in the pool. To enable the allele frequency of an SNP or allelic variant in the pool to be determined, several measurements need to be taken, to enable the allele frequency to be calculated. The height of the peak (see before) for each allelic variant residue needs to be measured, which should be present adjacent to each other on the pattern of nucleotide incorporation obtained. The calculation of allele frequency can thus be performed as follows:

$$\text{Allele frequency (Allele 2)} = \frac{\text{Peak Height (allele 2)}}{\text{Peak Height (allele 2)} + \text{Peak Height (allele 1)}} \times 100\%$$

Therefore, if the SNP is C/T the calculation would be performed thus:

$$\text{Allele frequency } T = \frac{\text{Peak height } T}{\text{Peak height } T + \text{Peak height } C} \times 100\%$$

Thus, it is possible to obtain accurate, cost-effective and rapid information on SNP allele frequencies in a population using nucleic acid pooling and primer-extension reactions, by monitoring nucleotide incorporation.

The method of the invention relies upon the knowledge of the location and potential variants of the SNP or allelic variant, together with further known sequence information (e.g. with known sequences of conserved/semi-conserved regions) from which to determine an appropriate primer binding site and design a complementary extension primer. Using the method of the invention, the allele frequency of any SNP or allelic variant may be determined, whether present in coding or non-coding regions.

The invention also comprises kits for carrying out the method of the invention. These will normally include one or more of the following components:

optionally primer(s) for in vitro amplification; a primer for the primer extension reaction; nucleotides for amplification and/or for the primer extension reaction (as described above); a polymerase enzyme for the amplification and/or primer extension reaction; and means for detecting primer extension (e.g. means of detecting the release of pyrophosphate as outlined and defined above).

The invention will now be described by way of nonlimiting examples.

EXAMPLE 1

Templates and Primers

These examples used DNA from 3 different sources which was either extracted from cell lines or from genomic sources. In total, DNA from 122 individual sources was used. The concentration of nucleic acid in some of the samples had been determined previously by measurement of absorbance at a wavelength of 260 nm. These samples were diluted to 2 ng/μl based on the absorbance measurements and the samples were either pooled directly, or after concentration calibration.

Some examples were performed on template oligonucleotides instead of PCR products. These oligonucleotides were obtained from Interactiva Ulm, Germany.

PCR amplification primers and sequencing primers were designed using Oligo 6.0 (Med Probe AS, Oslo, Norway). All primers were ordered from Interactiva (Supra).

TABLE 1

Primers and SNP definitions

SNP_ID	Upstream primer	Downstream primer	Sequencing primer	Fragment length [bp]	Sequencing output
Eu1 (ACP-240)	E1a (SEQ ID NO: 3) 5'-Biotin-ggt cgg gct ggg aag at-3'	E1b (SEQ ID NO: 4) 5'-gct ccc gca gag gaa gc-3'	E1s (SEQ ID NO: 5) 5'-aga aag ggc ctc ctc tct tt-3'	158	A/T
Eu4 (ACEex 15)	E4a (SEQ ID NO: 6) 5'-gcc agg aag ttt gat gtg aac-3'	E4b (SEQ ID NO: 7) 5'-Biotin-gat tcc cct ctc cct gta cct-3'	E4s (SEQ ID NO: 8) 5'-gac cta gaa cgg gca gc 3'	145	A/G
Eu7 (ANP1218)	E7a (SEQ ID NO: 9) 5'-Biotin-tga tgt aac cct cct ctc ca3'	E7b (SEQ ID NO: 10) 5'-cgg ctt acc ttc tgc tgt agt-3'	E7s (SEQ ID NO: 11) 5'-acg gca gct tct tcc cc-3'	142	C/T
460R	PSO 145 (SEQ ID NO: 12) 5'-B-ggc tgc tgt tct gaa acc atc tga -3'	PSO 146 (SEQ ID NO: 13) 5'-ttc agg aac gcg ggc aag tc -3'	PSO 147 (SEQ ID NO: 14) 5'-gag cag tcc cca ccc -3'	101	CC/T
461R	Same as 460R	Same as 460R	PSO 148 (SEQ ID NO: 15) 5'-gcg ggc aag tcc aat -3'	Same as 460R	C/TT
465R	PSO 149 (SEQ ID NO: 16) 5'-B-gga aca ctg cct ccc act ttc tt-3'	PSO 150 (SEQ ID NO: 17) 5'-tcc cca tgc agc cct aga gac-3'	PSO 151 (SEQ ID NO: 18) 5'-gga gaa gtc cag tgt gc -3'	85	C/T
466F	PSO 182 (SEQ ID NO: 19) 5'-ttc caa agg acg cga cca taa-3	PSO 183 (SEQ ID NO: 20) 5'-B-cct gca ccc cag acc act ga-3'	PSO 184 (SEQ ID NO: 21) 5'-tag ctg cgc ggg aa -3'	111	C/T/G
470R	PSO 155 (SEQ ID NO: 22) 5'-B-cct acc cac agg cca gaa-3'	PSO 156 (SEQ ID NO: 23) 5'-gcc tgg gac ctc act gtc -3'	PSO 157 (SEQ ID NO: 24) 5'-gga gac aga atg ctg at -3'	102	C/A
471F	PSO 158 (SEQ ID NO: 25) 5'-gtt gcc ctc tgg ttc cac ct -3'	PSO 159 (SEQ ID NO: 26) 5'-B-tgt ctc cag cag ctc ctt cat c -3'	PSO 160 (SEQ ID NO: 27) 5'-gcc cag gaa gga ac -3'	126	CCC/T

TABLE 1-continued

Primers and SNP definitions

SNP_ID	Upstream primer	Downstream primer	Sequencing primer	Fragment length [bp]	Sequencing output
481R	PSO 167 (SEQ ID NO: 28) 5'-B-gat gct gta aca gag acc cca ta -3'	PSO 168 (SEQ ID NO: 29) 5'-ctg gga tta cag gtg tga aca ct -3'	PSO 169 (SEQ ID NO: 30) 5'-tag gag caa gaa gta aac -3'	110	T/G
486R	PSO 173 (SEQ ID NO: 31) 5'-B-caa ggt aga gaa gtg cag cat tca -3'	P50 174 (SEQ ID NO: 32) 5'-ttg att ctc ttt gag ccc aga tgt -3'	PSO 175 (SEQ ID NO: 33) 5'-gcc tgg agc tgt taa t -3'	115	TT/C
1000F	PSO 194	PSO 195	PSO 196	159	CC/T
3345F	PSO 199	PSO 200	PSO 201	120	A/GGGG

TABLE 2

Oligonucleotides used to create "artificial" SNPs.

SNP name	Oligoname	Oligo Sequence	Sequencing output
Oligo 1	PSO43SNP	AGTCATGGTGCT**G**GGGCACTG **GCCGTCGTTTTACAACG** (SEQ ID NO: 34)	CCCC/T
	PSO44SNP	AGTCATGGTGCT**A**GGGCAGTG **GCCGTCGTTTTTACAACG** (SEQ ID NO: 35)	
Oligo 2	PSO44SNP	AGTCATGGTGCT**G**GGGGCACT **GGCCGTCGTTTTACAACG** (SEQ ID NO: 36)	CCCCC/T
	PSO45SNP	AGTCATGGTGCT**A**GGGGCACT **GGCCGTCGTTTTACAACG** (SEQ ID NO: 37)	
Oligo 3	PSO53SNP	AGTCATGGTGCTAA**G**GGGGCA CTG**GCCGTCGTTTTACAACG** (SEQ ID NO: 38)	CCCCC/ TTT
	PSO54SNP	AGTCATGGTGCTAA**A**GGGGCA CTG**GCCGTCGTTTTACAACG** (SEQ ID NO: 39)	
Sequencing primer	PSO55NUSPT	CGT TGT AAA ACG ACG GC (SEQ ID NO: 40)	

PCR Amplification

All fragments in the examples were amplified with the AmpliTaq Gold Kit (Applied Biosystems) and 2 mm $MgCl_2$, according to the following protocol:

PCR mix	1 x mix [µl]
GeneAmp 10xPCR butter II	5
$MgCl_2$ (25 mM)	4
DNTP (2.5 mM)	2.5
DMSO	0
Primer a (10 µM)	1
Primer b (10 µM)	1
TaqGold (5 U/µl)	0.3
H_2O	31.2
Sum	45

Approximately 10 ng genomic DNA was added to 45 µl of PCR mix to make a total PCR volume of 50 µl. The PCR cycling conditions were as follows: 95 C for 5 minutes, 45 cycles of (95 C for 15 seconds, Ta C for 30 seconds, 72 C for 15 seconds), 72 C for 5 minutes, 4 C. For SNPs Eu1, Eu4 and Eu7 Ta=57 C. Otherwise Ta=60 C.

EXAMPLE 2

DNA Calibration

In order to calibrate the amount of DNA in each of the samples, an SNP was chosen for analysis. SNP 465R was chosen, it is a C/T SNP that generates good signals without preferential amplification, is not present in a homopolymeric stretch and gives no background signals or uneven peak heights. All samples were genotyped for the chosen SNP.

TABLE 3

Primers used to amplify and sequence SNP 465R.

SNP ID	Upstream primer	Downstream primer	Sequencing primer length	Fragment SNP	Sequencing output
465R	5-B-gga aca ctg cct ccc act ttc tt-3' (SEQ ID NO: 16)	5-tcc cca tgc agc cct aga gac-3 (SEQ ID NO: 17)	5-gga gaa gtc cag tgt gc-3 (SEQ ID NO: 18)	85	G/AC/T

The genotyping was performed as follows. 5 μl genomic DNA (at a concentration of approximately 2 ng/μl) was amplified as described previously in Example 1. 25 μl of the PCR product was mixed with 8 μl magnetic beads Dynabeads® (Dynal Biotech ASA, Oslo, Norway) (10 μg/μl) and 17 μl 2×BW buffer (10 mm Tris-HCl, 2M NaCl, 1 mM EDTA, 0.1% Tween 20). The strands were then separated using 50 μl 0.5M NaOH. The sample was then treated with 1× annealing buffer (20 mM Tris-acetate, 5 mM MgAc), and washed. The beads were transferred to a PSQ 96™ plate (Pyrosequencing AB, Uppsala, Sweden) which contained 40 μl of 1× annealing buffer and 5 μl sequencing primer. A sequencing reaction was then performed on a PSQ 96™ instrument (Pyrosequencing AB) using SNP reagent kit, product number 40-0001 (Pyrosequencing AB). Once the genotype of SNP 465R of each sample had been established, calibration was performed.

2.5 μl of sample genomic DNA (at an approximate concentration of 2 ng/μl) was added to 2.5 μl reference genomic DNA and 45 μl PCR mix added, and PCR performed (supra).

The SNP was then analysed (as for genotyping assay) on a PSQ 96™ instrument (Pyrosequencing AB) using Pyrosequencing™ reagents (product no 40-0001).

Calculations and data:

Reference #1: T/T

Reference #2: C/C

Conc (Reference #2)=$X_{Ref\ \#2}$× Conc (Reference #1)

Conc (sample)=X×Conc (Reference #1)

Calculation of $X_{REf\ \#2}$ and $Y_{Ref\ \#2}$:

Reference #2+Reference #1 are pooled:

$$X_{Ref\ \#2} = \frac{\text{Peak height } C}{\text{Peak height } T}$$

$$Y_{Ref\ \#2} = \frac{\text{Peak height } C}{(\text{Peak height } T + \text{Peak height } C)}$$

Calculation of X and Y for all other samples:

Homozygotes C/C sample+Reference #1 are pooled:

$$X = \frac{\text{Peak height } C}{\text{Peak height } T}$$

$$Y = \frac{\text{Peak height } C}{(\text{Peak height } T + \text{Peak height } C)}$$

Homozygote T/T sample+Reference #2 are pooled:

$$X = X_{Ref\ \#2} = \frac{\text{Peak height } T}{\text{Peak height } C} \quad Y = \frac{\text{Peak height } T}{(\text{Peak height } T + \text{Peak height } C)}$$

Heterozygote C/T+Reference #1:

$$X = \frac{2 \times \text{Peak height } C}{(\text{Peak height } T - \text{Peak height } C)}$$

$$Y = \frac{\text{Peak height } C}{(\text{Peak height } T + \text{Peak height } C)}$$

TABLE 4

Results for some of the calibrated samples.

Sample	Sample Genotype	Sample mix	Allele	Peak height	Y	X
Ref #2	C/C	ref #2 + ref #1	C	26.25	0.51	1.0
			T	25.62		
#1	C/C	#1 + ref #1	C	19.68	0.40	0.7
			T	30.07		
#2	C/T	#2 + ref #1	C	12.65	0.24	0.9
			T	41.09		
#3	C/T	#3 + ref #1	C	12.64	0.24	1.0
			T	39.09		
#18	T/T	#18 + ref #2	C	28.05	0.45	0.8
			T	23.05		
#19	T/T	#19 + ref #2	C	33.78	0.35	0.5
			T	18.13		

Thus, for further experiments, a given volume of reference #1 is put into the pool, and the X and Y values obtained for the samples can be used to determine the volume of each sample to be added to the pool.

$$\text{Volume (Sample \#1)} = \frac{\text{Volume } (Ref\ \#1)}{X\ \text{(Sample \#1)}}$$

$$\text{Volume (Sample \#19)} = \frac{\text{Volume } (Ref\ \#1)}{X\ \text{(Sample \#19)}}$$

TABLE 5

Calculated X and Y values and thus volume of sample to use in pooling nucleic acid samples

Sample	Sample Genotype	Sample mix	Allele	Peak height	Y	X	Volume (μl)
Ref #1	C/C	—	C	—	—	1.00	50
			T	—	—		
Ref #2	C/C	ref #2 + ref #1	C	26.25	0.51	1.02	49
			T	25.62			
#1	C/C	#1 + ref #1	C	19.68	0.40	0.65	76
			T	30.07			
#2	C/T	#2 + ref #1	C	12.65	0.24	0.90	56
			T	41.09			
#3	C/T	#3 + ref #1	C	12.64	0.24	0.96	52
			T	39.09			

TABLE 5-continued

Calculated X and Y values and thus volume of sample to use in pooling nucleic acid samples

Sample	Sample Genotype	Sample mix	Allele	Peak height	Y	X	Volume (µl)
#18	T/T	#18 + ref #2	C	28.05	0.45	0.84	59
			T	23.05			
#19	T/T	#19 + ref #2	C	33.78	0.35	0.55	91
			T	18.13			

Assessing DNA Calibration

20 samples were chosen. The DNA concentrations had been determined by using UV absorbance measurements and diluted to a concentration of 2 ng/µl. The 20 samples had been individually genotyped for the SNP (465R) using PSQ™ 96 system. The samples were pooled individually with a "reference DNA", also from the diversity panel. PCR was performed to amplify the fragment containing SNP 465R, and sequencing was performed on PSQ™ 96 system. The concentrations were compared with each other by calculations on the peak heights, and are tabulated in Table 6, below. Further, two test pools were made (one constructed using the calibrated concentrations (pool 1) and one using the original concentrations from UV absorbance measurements (pool 2).

TABLE 6

Calculations for DNA concentration adjustment

Sample	Sample Genotype	Sample mix	Allele	Peak height	Y	X	Z	Volume (µl)
Ref #2	C/C	ref #2 + ref #1	C	11,77	0,60	1,5	1,0	15
			T	7,79				
#1	C/T	#1 + ref #1	C	7,17	0,34	2,2	1,5	10
			T	13,63				
#2	C/T	#2 + ref #1	C	7,39	0,35	2,4	,16	9
			T	13,44				
#3	C/C	#3 + ref #1	C	11,42	0,60	1,5	1,0	15
			T	7,72				
#4	C/T	#4 + ref #1	C	6,77	0,37	2,9	1,9	8
			T	11,5				
#5	C/T	#5 + ref #1	C	8,4	0,41	4,5	3,0	5
			T	12,13				
#6	C/C	#6 + ref #1	C	9,02	0,52	1,1	0,7	21
			T	8,39				
#7	C/T	#7 + ref #1	C	8,14	0,38	3,0	2,0	7
			T	13,52				
#8	C/T	#8 + ref #1	C	8,47	0,42	5,2	3,5	4
			T	11,71				
#9	C/T	#9 + ref #1	C	8,02	0,39	3,5	2,3	6
			T	12,61				
#10	C/T	#10 + ref #1	C	6,71	0,29	1,4	0,9	16
			T	16,17				
#11	C/T	#11 + ref #1	C	6,25	0,30	1,5	1,0	15
			T	14,44				
#12	C/C	#12 + ref #1	C	14,2	0,66	1,9	1,3	12
			T	7,39				
#13	C/T	#13 + ref #1	C	7,84	0,37	2,9	1,9	8
			T	13,21				
#14	C/T	#14 + ref #1	C	6,67	0,36	2,7	1,8	8
			T	11,63				
#15	C/T	#15 + ref #1	C	3,08	0,20	0,7	0,4	34
			T	12,31				
#16	C/C	#16 + ref #1	C	11,82	0,56	1,3	0,8	18
			T	9,29				
#17	C/C	#17 + ref #1	C	15,91	0,73	2,7	1,8	8
			T	5,96				
#18	T/T	#18 + ref #2	C	12,91	0,42	0,7	0,7	21
			T	9,41				
#19	T/T	#19 + ref #2	C	11,52	0,44	0,8	0,8	19
			T	8,88				

According to previous calculations for SNP465R observed differences in DNA concentrations would not have had any detectable impact on the allele frequency measurement for 465R in these pools. Expected allele frequency for the T-allele was 40% in pool 1 and 41% in pool 2, which is an undetectable difference. Therefore, two further SNPs were selected to test the pools, SNP 461R and 470R. The difference between the two pools was expected to be 3% for both SNPs and that is a detectable difference.

For both pools, the estimated allele frequencies were in good accordance with what was expected, see FIG. **1** and Table 7. The experiment showed that it is possible to use Pyrosequencing™ as a method to calibrate DNA concentrations before pooling DNA. Further, the calibrated pool was more in accordance with the theoretical frequencies, as determined from individual genotypes (10% for 461R and 55% for 470R).

TABLE 7

Measured allele frequencies and STD for each pool compared to the theoretically calculated frequencies of the DNA pools.

	461R Pool 1	461R Pool 2	470R Pool 1	470R Pool 2
Replicate 1	8.5	5.9	64.7	56.9
Replicate 2	6.1	7.2	55.8	54.1
Replicate 3	6.6	8.1	59.3	58.1
Replicate 4	9.3	4.8	51.6	59.8
Replicate 5	8.3	3.5	55.3	56.5
Replicate 6	6.7	5.6	56.1	59.2
Replicate 7	10.2	4.7	54.3	62.8
Replicate 8	7.1	6.6	57.1	58.5
Replicate 9	6.6	6.3	55.2	54.7
Replicate 10	6.9	3.8	57.4	55.5
average	7.6	5.6	56.8	57.6
calculated STD	10.0	7.0	55.0	58.0
	1.3	1.3	3.5	2.5

Therefore, this method of sequencing can also be used reliably for the calibration of relative concentrations in a pool of nucleic acid. This has applications for all sequencing-by-synthesis protocols.

EXAMPLE 3

SNP Analysis Protocol

The pooled DNA (calibrated according to Example 2, or of known concentration) was added to 45 μl PCR mix (supra) and amplified as described previously. 25 μl of the PCR product was mixed with 8 μl magnetic beads—Dynabeads® (Dynal Biotech ASA, Oslo, Norway) (10 μg/μl) as described in Example 2. Annealing of the primer to the template DNA was performed with 15 pmol sequencing primer, for 2 minutes at 80 C. The samples were allowed to cool to room temperature and the primer extension reaction was performed on a PSQ™ 96 instrument (Pyrosequencing AB) using SNP reagent kit (Pyrosequencing AB). Once the peak height data was collected for the DNA pool, the allele frequency can be calculated as follows if the SNP is not present in a homopolymeric stretch:—

Allele frequency (Allele 2)=

$$\frac{\text{Peak Height (Allele 2)}}{\text{Peak Height (Allele 2)} + \text{Peak Height (Allele 1)}} \times 100\%$$

EXAMPLE 4

Pooling Strategies

It is important to determine whether it is more preferable to pool genomic DNA or PCR product, as experimental variance can be expected once PCR amplification of the genomic DNA has been performed. Thus, the SNP Eu7 (A/G) was investigated, by sequencing the SNP in reverse (T/C).

Ninety samples were individually genotyped for Eu7 and thereafter pooled either before or after PCR amplification, with five replicate reactions performed for each pool. The expected allele frequency is 27% G. The experiment was repeated in 3 subset populations (30–40 samples out of the 90) with lower allele frequencies (15% G, 10% G and 5% G, respectively).

Each replicate of a genomic DNA- or PCR-pool, 40 μl PCR product was incubated with 15 μl magnetic beads (10 μg/μl) and 25 μl 2×BW buffer. The resulting single-peak height levels were about 40–60 RLU. The theoretical allele frequency values (determined from the individual sample genotypes) in the four tested sample sets were 27% G, 15% G, 10% G, and 5% G respectively.

Pooling of PCR products resulted in good estimates of allele frequencies in all four pools (26%, 17%, 11%, and 7% respectively), and with low variance between replicate sequencing reactions. Pooling of genomic DNA resulted in accurate results (28%, 17%, 12%, and 6% respectively), but with slightly larger variation between replicate pools.

The experiment indicated that pooling of genomic DNA is possible with the same accuracy as can be obtained with pooled PCR products. However, the replicate PCR amplifications on the genomic DNA pool introduces additional experimental variance. Pooling of genomic DNA may therefore require testing more replicate pools to obtain the same accuracy as when pooling PCR products.

It can also be concluded that 5% of the G-allele could be reliably detected showing that even low allele frequencies are capable of measurement using the method of the invention.

FIG. **2***a* represents graphically the allele frequency results for 5 replicate PCR products on each of 4 pools. It can be seen that the estimated allele frequency (%) is in close correlation with the measured frequency. FIG. **2***b* shows graphically the allele frequency results for pooled genomic DNA, 5 replicate reaction per pool. Although the measured allele frequency is slightly more variable for the genomic DNA when compared to the PCR products, the calculated mean values were still in close agreement with the estimated frequency.

Pooling of Genomic DNA

Ninety samples were individually genotyped for five different SNPs. One A/G-SNP (Eu4), one tri-allelic SNP (466F), one simple C/T-SNP (465R), one C/T-SNP followed by a T (461R), and one A/C-SNP (470R). A pool containing ninety genomic DNA samples was created without calibration of the DNA concentrations and therefore differed slightly in individual DNA concentrations. For Eu4, five replicate PCR reactions were performed. For the other four SNPs, ten replicate PCR reactions were used. All PCR amplifications were performed with 10 ng genomic DNA as starting material in the PCR reaction. For Eu4, 40 μl PCR product was used for sequencing. For the other four SNP assays, 30 μl of each PCR product was used for Sequencing. The average allele frequencies and standard deviations were calculated.

Results on allele frequencies were calculated for five different SNPs, the results for which are tabulated below:

TABLE 8

Results from pooling experiments

SNP	Sequence	Expected Frequency	Measured Frequency
466F	[C/T/G]AAGGTTGTCCT (SEQ ID NO: 1)	C 38.1%	C 40.8%
		T 37.5%	T 32.1%
		G 24.4%	G 27.1%
465R	[C/T]GTTCCACCT (SEQ ID NO: 2)	C 64.4%	C 65.1%
		T 35.6%	T 34.9%
461R	[C/T]TGCAGA	C 92.2%	C 96.5%
		T 7.8%	T 3.5%
470R	T[C/A]TCTGG	C 28.9%	C 28.2%
		A 71.1%	C 71.8%
Eu4	[A/G]CTGCCT	G 56.7%	G 56.0%
		A 43.3%	A 44.0%

The sequencing results are shown as "pyrograms"™ (FIGS. **3***a*, **3***b*, **3***c*, **3***d* and **3***e*), wherein the peak height resulting from nucleotide addition is measured. No concentration calibration was performed for this experiment, and therefore different amounts of the individual nucleic acid samples were added to the pool. In view of this, the results are remarkably close to the estimated allele frequency for each pool. The standard deviation values for the results were between 0.8 and 1.8, which was found to be comparable with previous allele frequency experiments.

The result for the SNP 461R, which contains a T residue in a stretch of 2 T residues showed a lower value than expected. From further experimentation, this result turned out to be consistent for this allele, probably due to the fact that the SNP was present in a homopolymeric stretch.

The pyrogram™ for SNP Eu4 (FIG. **3***e*) shows very high and wide peaks. This was due to the use of 40 μl of PCR product.

Detecting Allele Frequency Differences Between Pools

Four sample pools, composed of 39–90 genomic DNA samples were constructed for both SNP 465R and SNP 461R. DNA concentration calibration was not performed before pooling. Allele frequencies were measured for 10 replicate reactions of each pool. 10 ng genomic DNA was used in a 50 μl PCR reaction and 30 μl of the PCR product was used for the primer extension reactions. The average allele frequencies and standard deviations were calculated. 95% and 99% confidence intervals were also estimated for the measured allele frequencies.

As previously observed, the measured frequencies for the T-allele of SNP 461R are too low. However, the deviation proved to be consistent, enabling detection of even small differences in allele frequencies between pools. The smallest sample pool, SNP465R:4 with 39 samples, showed the largest deviation from the expected frequency, indicating the importance and difficulty of DNA pool construction.

TABLE 9

Pool ID and % T calculated values

Pool ID	Pool Size (N)	% T
SNP465R:1	90	35.6
SNP465R:2	71	33.7
SNP465R:3	55	30.6
SNP465R:4	39	25.0
SNP461R:1	90	7.8
SNP461R:2	80	9.8
SNP461R:3	67	12.8
SNP461R:4	58	17.8

TABLE 10

Results for SNP456R and SNP46lR

Pool ID	% T	Std[%]	% T [95% Conf. Interval]	% T [99% Conf. Interval]
SNP465R:1	34.9	0.9	34.3–35.5	34.0–35.8
SNP465R:2	31.6	1.4	30.6–32.6	30.2–33.0
SNP465R:3	28.6	0.7	28.1–29.1	27.9–29.3
SNP465R:4	27.3	1.4	26.3–28.3	25.9–28.7
SNP461R:1	3.5	1.2	2.6–4.4	2.3–4.7
SNP461R:2	6.1	0.9	5.5–6.7	5.2–7.0
SNP461R:3	8.6	1.6	7.5–9.7	7.0–10.2
SNP461R:4	15.4	1.3	14.5–16.3	14.1–16.7

EXAMPLE 5

Peak Height Linearity

To establish that a correlation exists between peak heights obtained in a primer-extension reaction, and the underlying allele frequency, 3 SNPs were investigated, Eu1, Eu4 and Eu7. The DNA samples were amplified according to Example 1. Following PCR amplification, 2 homozygote samples were mixed in proportions in 5% increments from 0% to 100% (i.e. 0:100, 5:95, . . . , 100:0). The primer-extension reaction was performed according to Example 3, and the allele frequencies calculated. 5 pmol PCR product was used for each primer-extension reaction, resulting in single peak height levels that were about 30–40 RLU (relative light units). The peak heights in RLU were plotted against the expected allele frequencies (FIGS. **4***a*, **4***b* and **4***c*). A linear relationship over the complete range of tested allele frequencies was confirmed. Thus, the correlation between primer-extension peak heights and SNP allele frequencies is excellent. FIG. **5** depicts the linear relationship between allele frequency and peak height, and shows the peak height results for 3 primer extension reactions: 25% C, 50% C and 75% C.

SNPs Present in Homopolymeric Stretches

To establish whether the presence of a homopolymeric stretch over an SNP alters the applicability of the method of the invention, primer-extension reactions were performed for 3 SNPs. Synthesized oligonucleotides (Interactiva, supra) were used in order to obtain an SNP where both alleles are located in a homopolymer, or where the SNP lies in a homopolymer of 3 or more identical residues.

Prior to all experiments, the DNA pools were calibrated using the method described in Example 2. For each SNP, 10 replicates of individual genotypes were analyzed in order to obtain reference data for comparison with the pools. The following SNPs were investigated:

1000F is a C/T-SNP which is preceded by a C. 24 samples were used to create five pools with different expected allele frequencies. (3,8% C, 7,1% C, 10% C, 31,2% C and 39,4% C). In the experiment, ten replicates were analyzed for each pool.

345F is an A/G-SNP followed by GGG. 24 samples were used to create two pools with an expected allele frequency of 26% A and 10% A respectively. Both pools were sequenced with two different dispensation orders to achieve either two or three peaks for the SNP. In the experiment, ten replicates were analyzed for each pool.

SNP471F is a C/T SNP preceded by CC. Eight samples were used to create four different pools with an expected allele frequency of 4.5% T, 8% T, 21% T and 31% T respectively. In the experiment, ten replicates were analyzed for each pool.

Oligo 1, Oligo 2 and Oligo 3 are artificially created SNPs that were made by mixing two oligonucleotides that only differ in one base. (See table 2). The two differing oligonucleotides were in each case mixed together with the following ratios: 5:95, 10:90, 20:80, 50:50, 80:20, 90:10 and 95:5. Oligo 1 is a C/T SNP preceded by CCC, Oligo 2 is a C/T SNP preceded by CCCC, and Oligo 3 is a C/T SNP preceded by CCCC and followed by TT.

Results

1. SNP 1000F. (CC/T)

Prior to the experiment this SNP was also used to calibrate the samples for the DNA pools. 30 μl of PCR product was incubated with 10 μl magnetic beads and 20 μl 2×BW-buffer. Pool 1 and Pool 2 show the difference in allele frequency between a calibrated pool (Pool 2) and a pool where the same volume of each sample has been used (Pool 1). Before the calibration, Pool 1 was expected to have an allele frequency of 31.2. This was based on the assumption that all samples were of the same DNA concentration. The calibration shows that this is not the case and based on the relative concentrations of the samples it is now possible to recalculate the expected allele frequency of Pool 1 to be 39.4, which is much closer to the allele frequency that was obtained in the experiment. The results for these experiments are represented graphically as FIG. **6**.

TABLE 11

The obtained allele frequencies for 1000F compared to the expected frequencies and the STD for each pool.

Replicate	Pool 1	Pool 2	Pool 3	Pool 4	Pool 5
1	40.9	31.5	12.2	11.3	9.1
2	43.4	35.2	14.8	12.3	9.9
3	43.6	34.1	14.1	13.0	8.8
4	42.0	35.9	14.0	11.9	8.9
5	42.2	37.4	14.8	11.9	8.9
6	43.1	34.3	11.3	12.8	8.7
7	43.4	36.1	13.1	11.7	7.3
8	45.1	32.7	13.0	12.5	7.4
9	39.1	34.0	14.3	12.5	9.3
10	46.6	33.4	13.6	9.3	8.9
average	42.9	34.4	13.5	11.9	8.7
expected	39.4	34.2	10	7.1	3.8
STD	2	1.66	1.09	1	0.76

2. SNP 345F (A/GGGG).

30 μl of PCR product was incubated with 10 μl of magnetic beads and 20 μl of 2×BW-buffer. Two pools were made with the expected allele frequencies of 10% A and 26% A.

A comparison was made between a dispensation order (i.e. order of addition of nucleotides in the primer extension reaction) that generates two peaks and one that generates three peaks if the sample is a heterozygote. The small differences in allele frequency between the two different dispensation orders indicates that the result is not significantly influenced by whether the SNP has two or three informative peaks. The results are depicted graphically as FIGS. **8***a* and **8***b*.

In this SNP the A-peak reduction factor was set to 80% due to the higher peak obtained when using modified dATP (dATP S). This was based on calculations of allele frequencies in a run with individual samples. (The individual samples were run with a dispensation order that generates three peaks.) Despite with adjustment the SNP does not show a completely linear relationship between peak heights and allele frequency for individual samples. The obtained pool results are higher than expected, with the largest aberration in the lower frequencies. If the pool results are compared with the frequencies for 345F in individual samples (FIG. **7**) one can see that the pattern is similar. However, it is difficult to do any allele frequency studies on a SNP that is not linear. The results for this SNP are depicted graphically as FIG. **7**. The standard line shows an imaginary pattern for an "ideal" SNP.

TABLE 12

The obtained allele frequencies for 345 F. compared to the expected frequencies and the STD for each pool.

Replicate	Pool 1 2 peaks	Pool 1 3 peaks	Pool 2 2 peaks	Pool 2 3 peaks
1	36.0	35.7	14.5	15.5
2	35.8	33.7	17.2	17.2
3	34.5	34.6	13.6	16.3
4	36.6	35.2	15.2	15.8
5	33.2	32.9	11.4	12.4
6	34.1	35.1	12.2	13.9
7	33.7	35.0	12.7	15.4
8	32.8	35.5	12.5	16.1
9	35.7	31.2	14.4	16.8
10	34.0	33.7	13.6	15.6
average	34.6	34.3	13.7	15.5
expected	26	26	10	10
STD	1.23	1.33	1.6	1.35

3. SNP471F (CCC/T).

30 μl of PCR product was incubated with 101 μl of magnetic beads and 20 ul 2×BW-buffer. Four pools were made with the expanded allele frequencies of 68.7% C, 78.6% C, 91.7% C and 95.5% C.

TABLE 13

The obtained allele frequencies for SNP471F compared to the expected frequencies and the STD for each pool. The results are depicted graphically as FIG. 9. The standard line shows an imaginary pattern for an "ideal" SNP.

Replicate	Pool 1	Pool 2	Pool 3	Pool 4
1	64.0	76.6	87.6	93.1
2	61.2	73.3	86.1	91.7
3	62.3	76.9	86.0	92.0
4	66.0	76.7	86.7	91.0
5	65.3	79.8	85.5	91.9
6	57.5	77.3	86.3	90.0
7	68.6	79.3	85.6	90.1
8	68.0	78.2	84.3	92.0
9	70.5	74.5	88.2	90.7
10				91.1
average	64.8	77.0	86.2	91.5
expected	68.7	78.6	91.7	95.5
STD	3.83	1.96	1.1	0.81

4. Oligo 1 (CCCC/T), Oligo 2 (CCCCC/T) and Oligo 3 (CCCCC/TTT).

The two oligonucleotides used for each artificial SNP were mixed in different ratios to a final concentration of 1 pmol/µl. 2 ul of each mix were annealed with 10 pmol of sequencing primer in a volume of 45 µl.

The obtained average allele frequencies for Oligo 1 and 2 (FIG. **10***b*) are within 10% from the expected frequencies although the results do not seem to be completely linear. Oligo 3 (FIG. **10***c*) shows that a SNP with two homopolymeric stretches can not be expected to give reliable allele frequencies; it is far from the expected frequencies. A cumulative representation of the results is shown as FIG. **10***d*.

EXAMPLE 6

Template Quantity

It is important to use the correct amount of nucleic acid in order to reliably estimate allele frequency. To investigate the amount of genomic DNA required prior to amplification, the SNP465R was investigated. 10 ng, 1 ng, 0.1 ng and 0.05 ng DNA was added in 4 PCR amplification and subsequent primer-extension reactions. Four DNA pools were created from genomic DNA, with allele frequencies of 31% C, 19% C, 12.5% C and 6% C. Standard calibration was performed 20 µl of PCR product was used in primer-extension.

Results

The experiment showed a significant correlation between the amount DNA used in the PCR reaction and the variation between replicates. In samples where 10 ng DNA were used in the PCR, the deviations between replicates were small but increased quickly when the template amount was lowered. But even for samples where only 0.05 ng DNA were used, the average allele frequencies of 10 replicates were in good accordance with the expected. A template amount of at least 10 ng is required for a reliable allele frequency quantification if only one or few replicates are used. If many replicates are amplified, the average allele frequency will be correct even with lower DNA amount but the variation between replicates will be significant. The results are depicted graphically on FIGS. **11***a, b, c* and *d*.

Required Signal Level

The height of the peak measured during primer-extension is correlated to many factors, including the amount of PCR product used. In order to determine the threshold signal level to calculate allele frequencies, several experiments were performed. Four different SNPs with different expected allele frequencies were used. One C/A-SNP (470R), one T/G-SNP (481R), one T/C-SNP with a T before the SNP (486R) and one C/T-SNP with a C before the SNP (460R). For SNP 470, a pool was created of several genomic samples. The expected allele frequency was 55% A in this pool. For the other SNPs a different pool of samples was used. The expected allele frequencies in that pool was 19.5% G for SNP481R, 12.5% C for SNP486R and 6% G for SNP460R.

Results

The peak heights do not seem to affect the allele frequency results in any dramatic way. If the single peak height is below 10 RLU, the signal-to-noise ratio might be too low for the SNP, if one of the alleles is represented at a low frequency. Although quite small, the variation between replicate reactions seems to increase slightly when the average single-peak height level gets below 15 RLU. The results are represented graphically as figure (12).

All references cited herein are incorporated herein in their entirety.

SEQUENCE LISTING

<160> NUMBER OF SEQ ID NOS: 40

<210> SEQ ID NO 1
<211> LENGTH: 12
<212> TYPE: DNA
<213> ORGANISM: Homo sapiens
<220> FEATURE:
<221> NAME/KEY: misc_feature
<222> LOCATION: ()..()
<223> OTHER INFORMATION: n can be c, t or g

<400> SEQUENCE: 1

naaggttgtc ct 12

<210> SEQ ID NO 2
<211> LENGTH: 10
<212> TYPE: DNA
<213> ORGANISM: Homo sapiens
<220> FEATURE:
<221> NAME/KEY: misc_feature
<222> LOCATION: ()..()
<223> OTHER INFORMATION: n is c or t

<400> SEQUENCE: 2

ngttccacct 10

-continued

<210> SEQ ID NO 3
<211> LENGTH: 17
<212> TYPE: DNA
<213> ORGANISM: artificial sequence
<220> FEATURE:
<221> NAME/KEY: misc_feature
<222> LOCATION: ()..()
<223> OTHER INFORMATION: Primer - E1a

<400> SEQUENCE: 3

ggtcgggctg ggaagat 17

<210> SEQ ID NO 4
<211> LENGTH: 17
<212> TYPE: DNA
<213> ORGANISM: artificial sequence
<220> FEATURE:
<221> NAME/KEY: misc_feature
<222> LOCATION: ()..()
<223> OTHER INFORMATION: Primer -E1b

<400> SEQUENCE: 4

gctcccgcag aggaagc 17

<210> SEQ ID NO 5
<211> LENGTH: 20
<212> TYPE: DNA
<213> ORGANISM: artificial sequence
<220> FEATURE:
<221> NAME/KEY: misc_feature
<222> LOCATION: ()..()
<223> OTHER INFORMATION: Primer - E1s

<400> SEQUENCE: 5

agaaagggcc tcctctcttt 20

<210> SEQ ID NO 6
<211> LENGTH: 21
<212> TYPE: DNA
<213> ORGANISM: artificial sequence
<220> FEATURE:
<221> NAME/KEY: misc_feature
<222> LOCATION: ()..()
<223> OTHER INFORMATION: Primer - E4a

<400> SEQUENCE: 6

gccaggaagt ttgatgtgaa c 21

<210> SEQ ID NO 7
<211> LENGTH: 21
<212> TYPE: DNA
<213> ORGANISM: artificial sequence
<220> FEATURE:
<221> NAME/KEY: misc_feature
<222> LOCATION: ()..()
<223> OTHER INFORMATION: Primer - E4b

<400> SEQUENCE: 7

gattcccctc tccctgtacc t 21

<210> SEQ ID NO 8
<211> LENGTH: 17
<212> TYPE: DNA
<213> ORGANISM: artificial sequence
<220> FEATURE:
<221> NAME/KEY: misc_feature
<222> LOCATION: ()..()

-continued

<223> OTHER INFORMATION: Primer - E4s

<400> SEQUENCE: 8

gacctagaac gggcagc 17

<210> SEQ ID NO 9
<211> LENGTH: 20
<212> TYPE: DNA
<213> ORGANISM: artificial sequence
<220> FEATURE:
<221> NAME/KEY: misc_feature
<222> LOCATION: ()..()
<223> OTHER INFORMATION: Primer - E7a

<400> SEQUENCE: 9

tgatgtaacc ctcctctcca 20

<210> SEQ ID NO 10
<211> LENGTH: 21
<212> TYPE: DNA
<213> ORGANISM: artificial sequence
<220> FEATURE:
<221> NAME/KEY: misc_feature
<222> LOCATION: ()..()
<223> OTHER INFORMATION: Primer - E7b

<400> SEQUENCE: 10

cggcttacct tctgctgtag t 21

<210> SEQ ID NO 11
<211> LENGTH: 17
<212> TYPE: DNA
<213> ORGANISM: artificial sequence
<220> FEATURE:
<221> NAME/KEY: misc_feature
<222> LOCATION: ()..()
<223> OTHER INFORMATION: Primer - E7s

<400> SEQUENCE: 11

acggcagctt cttcccc 17

<210> SEQ ID NO 12
<211> LENGTH: 24
<212> TYPE: DNA
<213> ORGANISM: artificial sequence
<220> FEATURE:
<221> NAME/KEY: misc_feature
<222> LOCATION: ()..()
<223> OTHER INFORMATION: Primer - PSO 145

<400> SEQUENCE: 12

ggctgctgtt ctgaaaccat ctga 24

<210> SEQ ID NO 13
<211> LENGTH: 20
<212> TYPE: DNA
<213> ORGANISM: artificial sequence
<220> FEATURE:
<221> NAME/KEY: misc_feature
<222> LOCATION: ()..()
<223> OTHER INFORMATION: Primer - PSO 146

<400> SEQUENCE: 13

ttcaggaacg cgggcaagtc 20

<210> SEQ ID NO 14
<211> LENGTH: 15

-continued

<212> TYPE: DNA
<213> ORGANISM: artificial sequence
<220> FEATURE:
<221> NAME/KEY: misc_feature
<222> LOCATION: ()..()
<223> OTHER INFORMATION: Primer - PSO 147

<400> SEQUENCE: 14

gagcagtccc caccc 15

<210> SEQ ID NO 15
<211> LENGTH: 15
<212> TYPE: DNA
<213> ORGANISM: artificial sequence
<220> FEATURE:
<221> NAME/KEY: misc_feature
<222> LOCATION: ()..()
<223> OTHER INFORMATION: Primer - PSO 148

<400> SEQUENCE: 15

gcgggcaagt ccaat 15

<210> SEQ ID NO 16
<211> LENGTH: 23
<212> TYPE: DNA
<213> ORGANISM: artificial sequence
<220> FEATURE:
<221> NAME/KEY: misc_feature
<222> LOCATION: ()..()
<223> OTHER INFORMATION: Primer - PSO 149

<400> SEQUENCE: 16

ggaacactgc ctcccacttt ctt 23

<210> SEQ ID NO 17
<211> LENGTH: 21
<212> TYPE: DNA
<213> ORGANISM: artificial sequence
<220> FEATURE:
<221> NAME/KEY: misc_feature
<222> LOCATION: ()..()
<223> OTHER INFORMATION: Primer - PSO 150

<400> SEQUENCE: 17

tccccatgca gccctagaga c 21

<210> SEQ ID NO 18
<211> LENGTH: 17
<212> TYPE: DNA
<213> ORGANISM: artificial sequence
<220> FEATURE:
<221> NAME/KEY: misc_feature
<222> LOCATION: ()..()
<223> OTHER INFORMATION: Primer - PSO 151

<400> SEQUENCE: 18

ggagaagtcc agtgtgc 17

<210> SEQ ID NO 19
<211> LENGTH: 21
<212> TYPE: DNA
<213> ORGANISM: artificial sequence
<220> FEATURE:
<221> NAME/KEY: misc_feature
<222> LOCATION: ()..()
<223> OTHER INFORMATION: Primer - PSO 182

<400> SEQUENCE: 19

-continued

ttccaaagga cgcgaccata a 21

<210> SEQ ID NO 20
<211> LENGTH: 20
<212> TYPE: DNA
<213> ORGANISM: artificial sequence
<220> FEATURE:
<221> NAME/KEY: misc_feature
<222> LOCATION: ()..()
<223> OTHER INFORMATION: Primer - PSO 183

<400> SEQUENCE: 20

cctgcacccc agaccactga 20

<210> SEQ ID NO 21
<211> LENGTH: 14
<212> TYPE: DNA
<213> ORGANISM: artificial sequence
<220> FEATURE:
<221> NAME/KEY: misc_feature
<222> LOCATION: ()..()
<223> OTHER INFORMATION: Primer - PSO 184

<400> SEQUENCE: 21

tagctgcgcg ggaa 14

<210> SEQ ID NO 22
<211> LENGTH: 18
<212> TYPE: DNA
<213> ORGANISM: artificial sequence
<220> FEATURE:
<221> NAME/KEY: misc_feature
<222> LOCATION: ()..()
<223> OTHER INFORMATION: Primer - PSO 155

<400> SEQUENCE: 22

cctacccaca ggccagaa 18

<210> SEQ ID NO 23
<211> LENGTH: 18
<212> TYPE: DNA
<213> ORGANISM: artificial sequence
<220> FEATURE:
<221> NAME/KEY: misc_feature
<222> LOCATION: ()..()
<223> OTHER INFORMATION: Primer - PSO 156

<400> SEQUENCE: 23

gcctgggacc tcactgtc 18

<210> SEQ ID NO 24
<211> LENGTH: 17
<212> TYPE: DNA
<213> ORGANISM: artificial sequence
<220> FEATURE:
<221> NAME/KEY: misc_feature
<222> LOCATION: ()..()
<223> OTHER INFORMATION: Primer - PSO 157

<400> SEQUENCE: 24

ggagacagaa tgctgat 17

<210> SEQ ID NO 25
<211> LENGTH: 20
<212> TYPE: DNA
<213> ORGANISM: artificial sequence
<220> FEATURE:
<221> NAME/KEY: misc_feature

-continued

<222> LOCATION: ()..()
<223> OTHER INFORMATION: Primer - PSO 158

<400> SEQUENCE: 25

gttgccctct ggttccacct 20

<210> SEQ ID NO 26
<211> LENGTH: 22
<212> TYPE: DNA
<213> ORGANISM: artificial sequence
<220> FEATURE:
<221> NAME/KEY: misc_feature
<222> LOCATION: ()..()
<223> OTHER INFORMATION: Primer - PSO 159

<400> SEQUENCE: 26

tgtctccagc agctccttca tc 22

<210> SEQ ID NO 27
<211> LENGTH: 14
<212> TYPE: DNA
<213> ORGANISM: artificial sequence
<220> FEATURE:
<221> NAME/KEY: misc_feature
<222> LOCATION: ()..()
<223> OTHER INFORMATION: Primer - PSO 160

<400> SEQUENCE: 27

gcccaggaag gaac 14

<210> SEQ ID NO 28
<211> LENGTH: 23
<212> TYPE: DNA
<213> ORGANISM: artificial sequence
<220> FEATURE:
<221> NAME/KEY: misc_feature
<222> LOCATION: ()..()
<223> OTHER INFORMATION: Primer - PSO 167

<400> SEQUENCE: 28

gatgctgtaa cagagacccc ata 23

<210> SEQ ID NO 29
<211> LENGTH: 23
<212> TYPE: DNA
<213> ORGANISM: artificial sequence
<220> FEATURE:
<221> NAME/KEY: misc_feature
<222> LOCATION: ()..()
<223> OTHER INFORMATION: Primer - PSO 168

<400> SEQUENCE: 29

ctgggattac aggtgtgaac act 23

<210> SEQ ID NO 30
<211> LENGTH: 18
<212> TYPE: DNA
<213> ORGANISM: artificial sequence
<220> FEATURE:
<221> NAME/KEY: misc_feature
<222> LOCATION: ()..()
<223> OTHER INFORMATION: Primer - PSO 169

<400> SEQUENCE: 30

taggagcaag aagtaaac 18

<210> SEQ ID NO 31

-continued

<211> LENGTH: 24
<212> TYPE: DNA
<213> ORGANISM: artificial sequence
<220> FEATURE:
<221> NAME/KEY: misc_feature
<222> LOCATION: ()..()
<223> OTHER INFORMATION: Primer - PSO 173

<400> SEQUENCE: 31

caaggtagag aagtgcagca ttca 24

<210> SEQ ID NO 32
<211> LENGTH: 24
<212> TYPE: DNA
<213> ORGANISM: artificial sequence
<220> FEATURE:
<221> NAME/KEY: misc_feature
<222> LOCATION: ()..()
<223> OTHER INFORMATION: Primer - PSO 174

<400> SEQUENCE: 32

ttgattctct ttgagcccag atgt 24

<210> SEQ ID NO 33
<211> LENGTH: 16
<212> TYPE: DNA
<213> ORGANISM: artificial sequence
<220> FEATURE:
<221> NAME/KEY: misc_feature
<222> LOCATION: ()..()
<223> OTHER INFORMATION: Primer - PSO 175

<400> SEQUENCE: 33

gcctggagct gttaat 16

<210> SEQ ID NO 34
<211> LENGTH: 38
<212> TYPE: DNA
<213> ORGANISM: artificial sequence
<220> FEATURE:
<221> NAME/KEY: misc_feature
<222> LOCATION: ()..()
<223> OTHER INFORMATION: oligonucleotide - PSO43SNP

<400> SEQUENCE: 34

agtcatggtg ctggggcact ggccgtcgtt ttacaacg 38

<210> SEQ ID NO 35
<211> LENGTH: 38
<212> TYPE: DNA
<213> ORGANISM: artificial sequence
<220> FEATURE:
<221> NAME/KEY: misc_feature
<222> LOCATION: ()..()
<223> OTHER INFORMATION: oligonucleotide - PSO44SNP

<400> SEQUENCE: 35

agtcatggtg ctagggcact ggccgtcgtt ttacaacg 38

<210> SEQ ID NO 36
<211> LENGTH: 39
<212> TYPE: DNA
<213> ORGANISM: artificial sequence
<220> FEATURE:
<221> NAME/KEY: misc_feature
<222> LOCATION: ()..()
<223> OTHER INFORMATION: oligonucleotide - PSO44SNP

<400> SEQUENCE: 36

-continued

agtcatggtg ctgggggcac tggccgtcgt tttacaacg 39

<210> SEQ ID NO 37
<211> LENGTH: 39
<212> TYPE: DNA
<213> ORGANISM: artificial sequence
<220> FEATURE:
<221> NAME/KEY: misc_feature
<222> LOCATION: ()..()
<223> OTHER INFORMATION: oligonucleotide - PSO45SNP

<400> SEQUENCE: 37

agtcatggtg ctaggggcac tggccgtcgt tttacaacg 39

<210> SEQ ID NO 38
<211> LENGTH: 41
<212> TYPE: DNA
<213> ORGANISM: artificial sequence
<220> FEATURE:
<221> NAME/KEY: misc_feature
<222> LOCATION: ()..()
<223> OTHER INFORMATION: oligonucleotide - PSO53SNP

<400> SEQUENCE: 38

agtcatggtg ctaaaggggc actggccgtc gttttacaac g 41

<210> SEQ ID NO 39
<211> LENGTH: 41
<212> TYPE: DNA
<213> ORGANISM: artificial sequence
<220> FEATURE:
<221> NAME/KEY: misc_feature
<222> LOCATION: ()..()
<223> OTHER INFORMATION: oligonucleotide - PSO54SNP

<400> SEQUENCE: 39

agtcatggtg ctaaaggggc actggccgtc gttttacaac g 41

<210> SEQ ID NO 40
<211> LENGTH: 17
<212> TYPE: DNA
<213> ORGANISM: artificial sequence
<220> FEATURE:
<221> NAME/KEY: misc_feature
<222> LOCATION: ()..()
<223> OTHER INFORMATION: Primer- PSO55NUSPT

<400> SEQUENCE: 40

cgttgtaaaa cgacggc 17

The invention claimed is:

1. A method of determining the frequency of an allele in a population of nucleic acid molecules, said method comprising:

pooling the nucleic acid molecules of said population to provide a pooled nucleic acid sample; performing primer extension reactions in a reaction mixture comprising said pooled nucleic acid sample and a primer which binds at a predetermined site located in said nucleic acid molecules, wherein said site is substantially adjacent to a polymorphic position of interest in said allele, to provide primer extension products, and wherein the primer extension reaction is performed by sequentially adding non-chain terminating nucleotides to the reaction mixture and quantitatively determining the incorporation or non-incorporation of each nucleotide as each nucleotide is added by bioluminometrically detecting the release of pyrophosphate; obtaining a pattern of nucleotide incorporation in said primer extension products at the positions that correspond to said polymorphic position of interest; and determining the frequency of said allele from said pattern of nucleotide incorporation.

2. The method according to claim 1 wherein ELIDA detection enzymes are used to detect the release of pyrophosphate.

3. The method according to claim 2 wherein a nucleotide-degrading enzyme is included during the primer extension reaction.

4. The method according to claim 1 wherein the nucleic acid molecules are immobilized on a solid support.

5. The method according to claim 1 wherein the amount or concentration of the nucleic acid in each sample of the population which is pooled, is determined prior to pooling.

6. The method according to claim 5 wherein the concentration of the nucleic acid in each sample of the population is determined by a primer-extension reaction prior to pooling.

7. The method according to claim 6 wherein the volume of each nucleic acid in each sample to be pooled is adjusted in view of the amount or concentration of nucleic acid present such that the pooled sample contains substantially the same amount or concentration of each nucleic acid molecule in the population.

8. The method according to claim 7 wherein a particular polymorphism is selected as a reference polymorphism and said primer extension reaction used to determine the concentration of nucleic acid in said sample is specific for said reference polymorphism.

9. The method according to claim 8 wherein said polymorphism is chosen such that it gives no background signals in a primer-extension reaction and that the signals are even.

10. The method according to claim 8 wherein said polymorphism is not present in a homopolymeric sequence and will not be preferentially amplified in any PCR-type reactions.

11. The method according to claim 8 wherein a reference sample containing said polymorphism is selected as the main reference from one of the homozygotes of one of the alleles of said polymorphism (Ref 1) and another reference (Ref 2) containing said polymorphism is selected from the other homozygote, and the reference samples are pooled and primer extension reactions are performed, and the pattern of nucleotide incorporation determined to determine the relative concentration of each reference sample.

12. The method according to claim 11 wherein the sample nucleic acid molecules to be tested are pooled individually with the reference samples.

13. A method of determining the amount of an allele in a sample of nucleic acid molecules, said method comprising:

performing primer extension reactions in a reaction mixture comprising said nucleic acid molecules, using a primer which binds at a predetermined site located in at least one said molecule wherein said site is substantially adjacent to a polymorphic position of interest in said allele, to provide primer extension products, and wherein the primer extension reaction is performed by sequentially adding non-chain terminating nucleotides to the reaction mixture and quantitatively determining the incorporation or non-incorporation of each nucleotide as each nucleotide is added by bioluminometrically detecting the release of pyrophosphate; determining the type and number of nucleotides incorporated in said primer extension products at positions that correspond to the polymorphic position of interest, and determining the amount of occurrence of said allele in said sample by analyzing the type and number of nucleotides incorporated.

14. The method according to claim 13 wherein ELIDA detection enzymes are used to detect the release of pyrophosphate.

15. The method according to claim 14 wherein a nucleotide-degrading enzyme is included during the primer extension reaction.

16. The method according to claim 15 wherein the nucleic acid molecules are immobilized on a solid support.

* * * * *

US 20020086324A1

(19) **United States**

(12) **Patent Application Publication** (10) **Pub. No.: US 2002/0086324 A1**

Laird et al. (43) **Pub. Date: Jul. 4, 2002**

(54) **PROCESS FOR HIGH THROUGHPUT DNA METHYLATION ANALYSIS**

(76) Inventors: **Peter W. Laird**, South Pasadena, CA (US); **Cindy A. Eads**, South Pasadena, CA (US); **Kathleen D. Danenberg**, Altadena, CA (US)

Correspondence Address:
DAVIS WRIGHT TREMAINE, LLP
2600 CENTURY SQUARE
1501 FOURTH AVENUE
SEATTLE, WA 98101-1688 (US)

(21) Appl. No.: **10/016,505**

(22) Filed: **Dec. 10, 2001**

Related U.S. Application Data

(63) Continuation of application No. 09/311,912, filed on May 14, 1999, now patented.

Publication Classification

(51) **Int. Cl.**[7] **C12Q 1/68**; C12P 19/34

(52) **U.S. Cl.** **435/6**; 435/91.2

(57) **ABSTRACT**

There is disclosed an improved high-throughput and quantitative process for determining methylation patterns in genomic DNA samples based on amplifying modified nucleic acid, and detecting methylated nucleic acid based on amplification-dependent displacement of specifically annealed hybridization probes. Specifically, the inventive process provides for treating genomic DNA samples with sodium bisulfite to create methylation-dependent sequence differences, followed by detection with fluorescence-based quantitative PCR techniques. The process is particularly well suited for the rapid analysis of a large number of nucleic acid samples, such as those from collections of tumor tissues

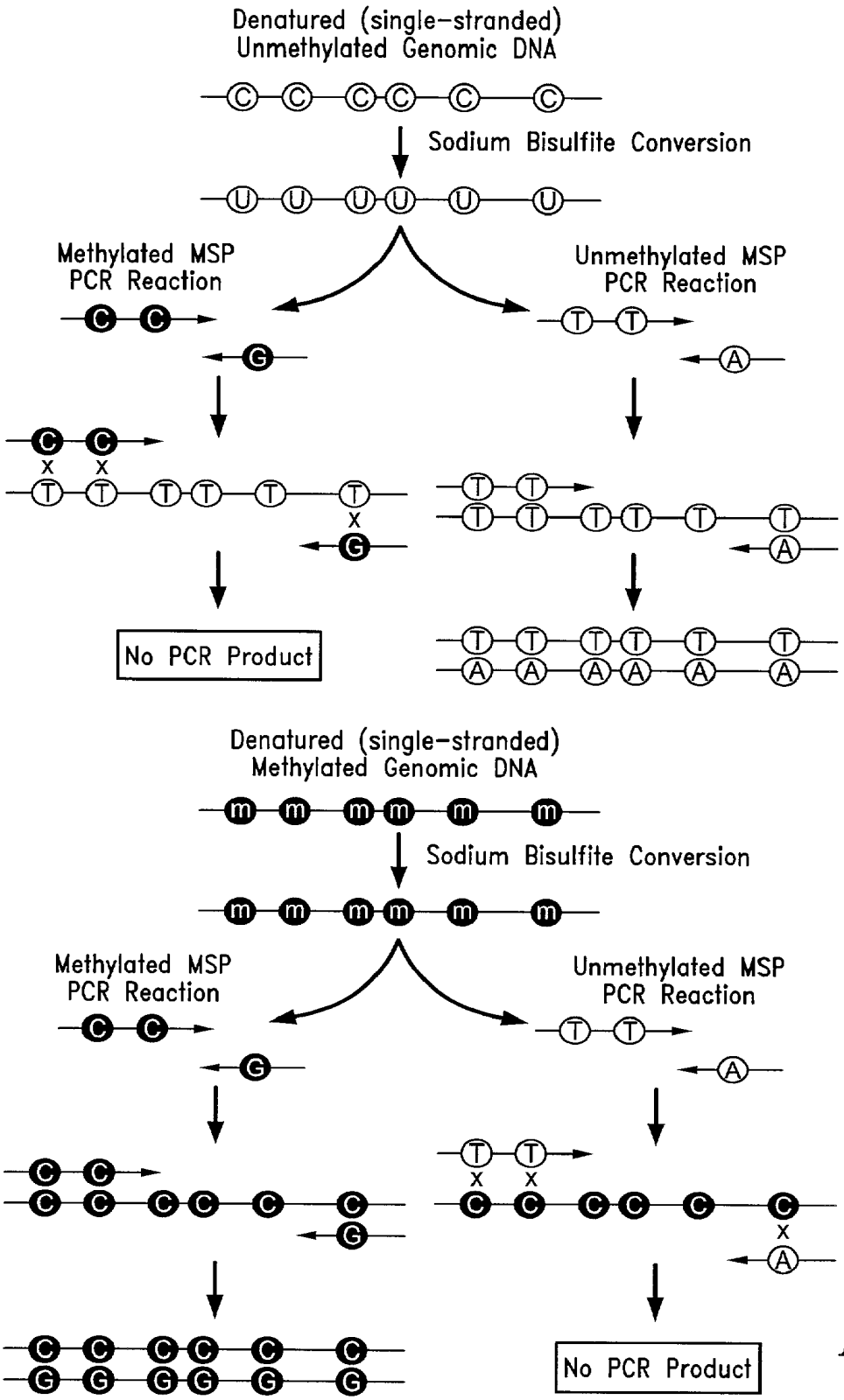

Denatured (single-stranded)
Unmethylated Genomic DNA
Sodium Bisulfite Conversion
Methylated MSP
PCR Reaction
Unmethylated MSP
PCR Reaction
No PCR Product
Denatured (single-stranded)
Methylated Genomic DNA
Sodium Bisulfite Conversion
Methylated MSP
PCR Reaction
Unmethylated MSP
PCR Reaction
No PCR Product

Fig. 1

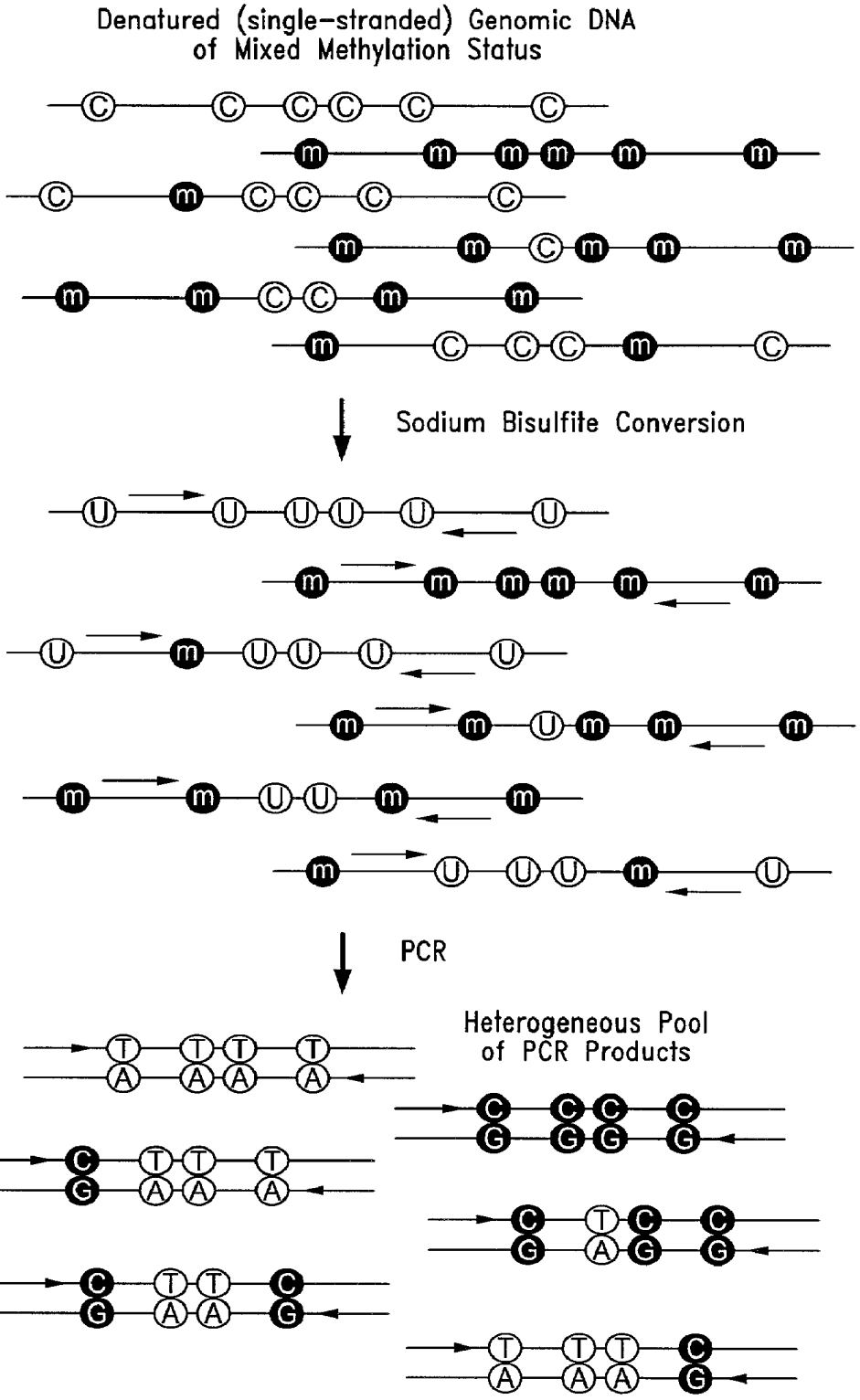

Denatured (single-stranded) Genomic DNA
of Mixed Methylation Status
Sodium Bisulfite Conversion
PCR
Heterogeneous Pool
of PCR Products

Fig. 2

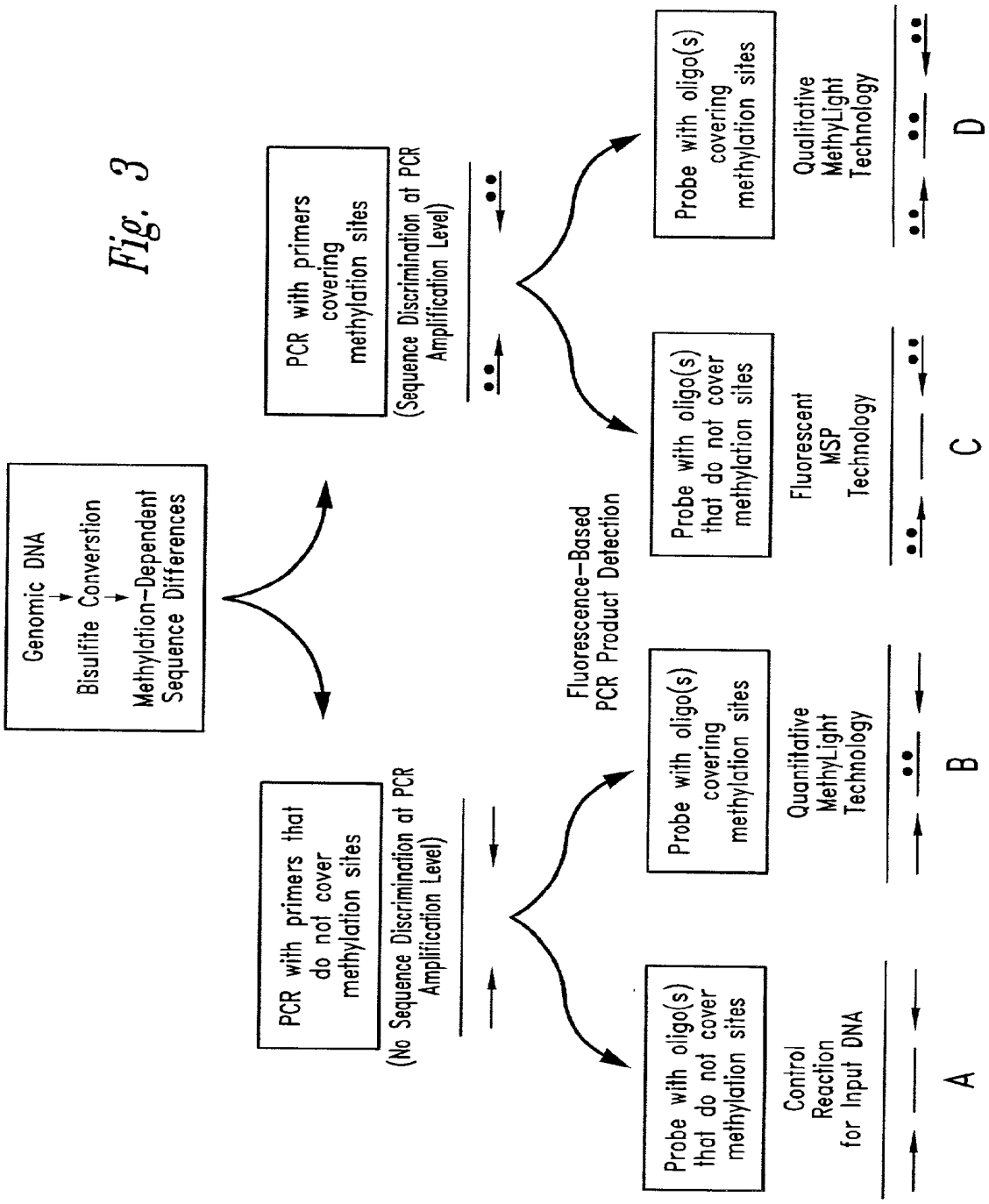

Genomic DNA
Bisulfite Converstion
Methylation-Dependent Sequence Differences
PCR with primers that do not cover methylation sites
(No Sequence Discrimination at PCR Amplification Level)
PCR with primers covering methylation sites
(Sequence Discrimination at PCR Amplification Level)
Fluorescence-Based PCR Product Detection
Probe with oligo(s) that do not cover methylation sites
Control Reaction for Input DNA
A
Probe with oligo(s) covering methylation sites
Quantitative MethyLight Technology
B
Probe with oligo(s) that do not cover methylation sites
Fluorescent MSP Technology
C
Probe with oligo(s) covering methylation sites
Qualitative MethyLight Technology
D

Fig. 3

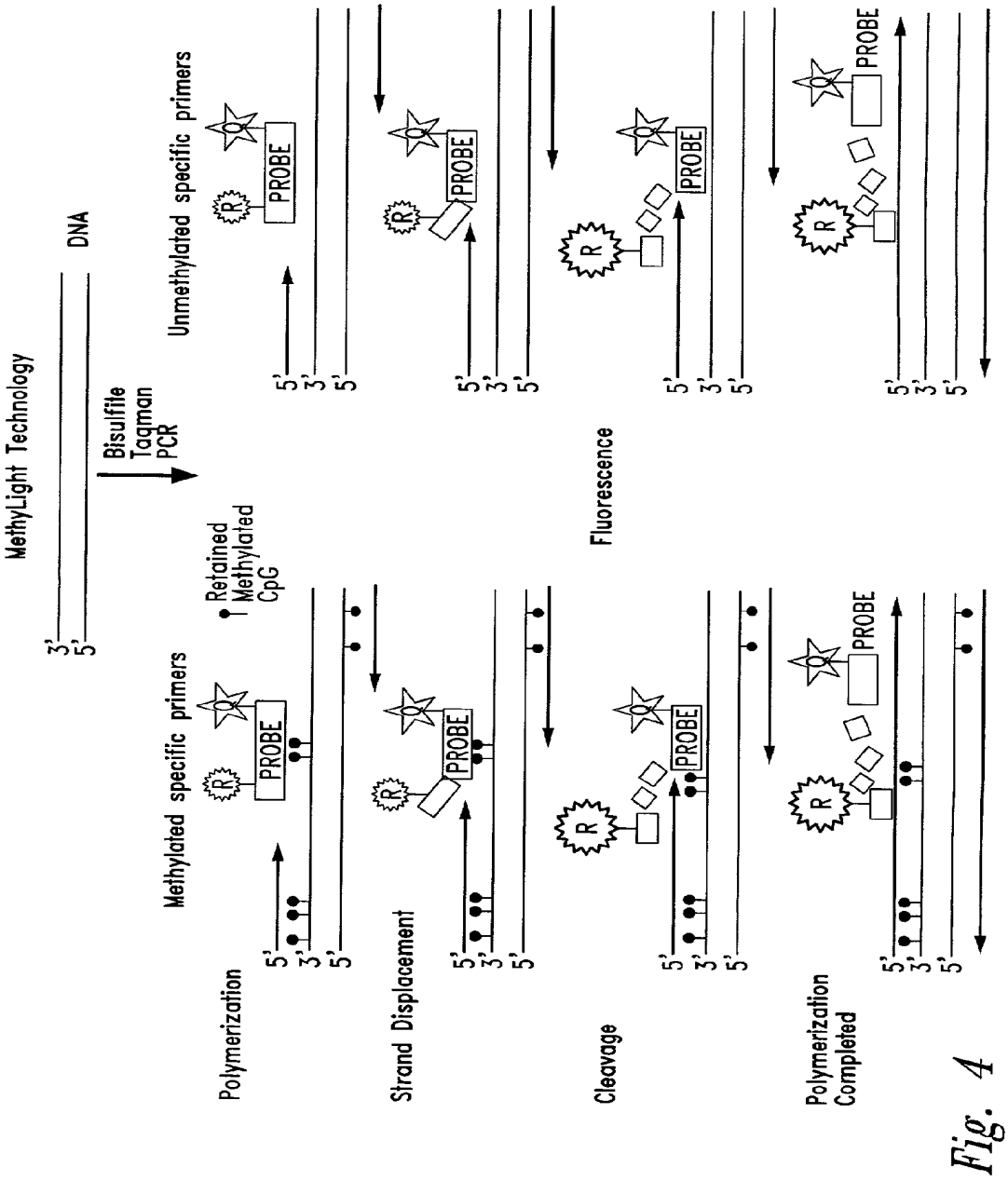

MethyLight Technology
DNA
3'
5'
Bisulfite
Taqman
PCR
Unmethylated specific primers
Methylated specific primers
Retained
Methylated
CpG
PROBE
R
5'
3'
Polymerization
Strand Displacement
Fluorescence
Cleavage
Polymerization
Completed

Fig. 4

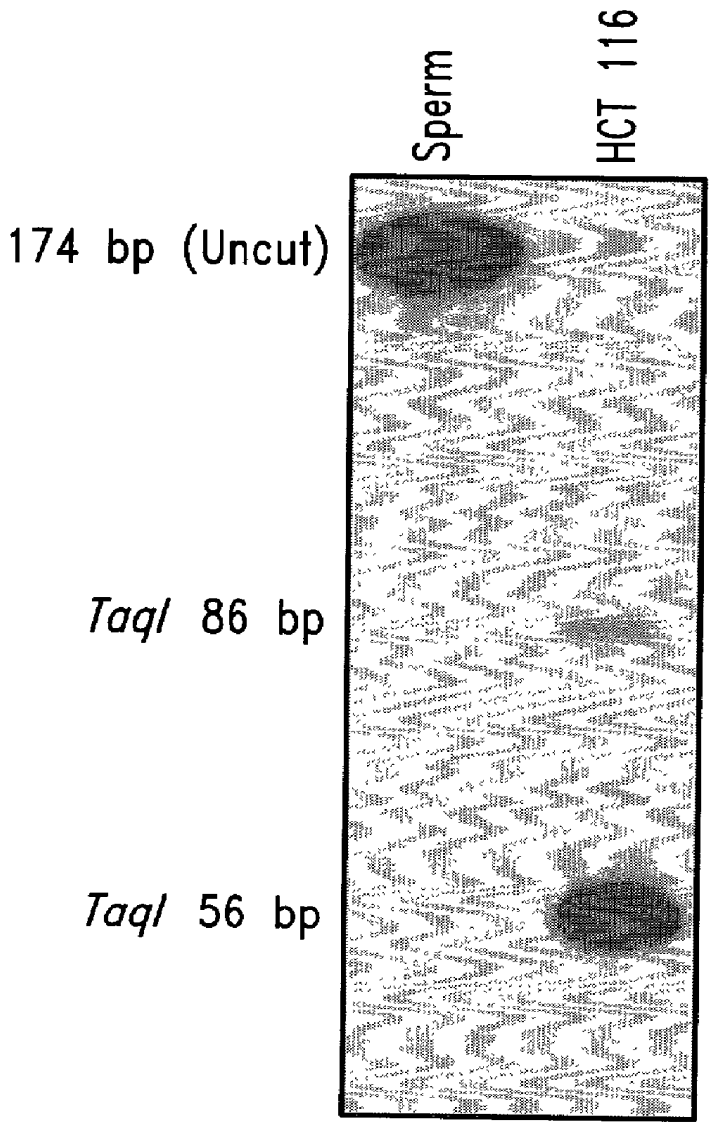

Sperm
HCT 116
174 bp (Uncut)
TaqI 86 bp
TaqI 56 bp
(86 bp) 0% 7%
(56 bp) 0% 88%
COBRA Gel

Fig. 5A

Methylation Analysis of the *ESR1* Gene

DNA	COBRA (% One or Two Sites Methylated)	MethyLight (Ratio of Methylated / Control)	MethyLight (Ratio of Unmethylated / Control)
Bisulfite Treated			
Sperm	0%	0	62
HCT116	95%	36	0
Untreated			
Sperm	No PCR	0	0
HCT116	No PCR	0	0

Fig. 5B

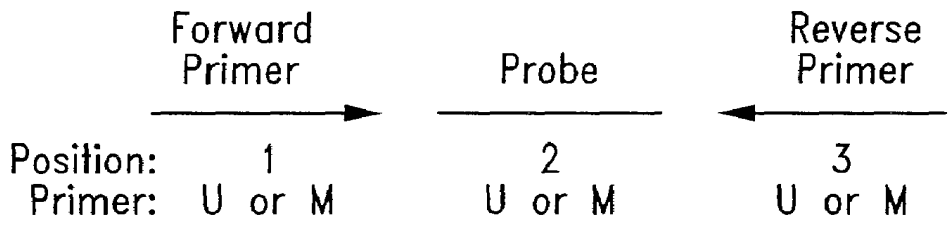


Fig. 6A

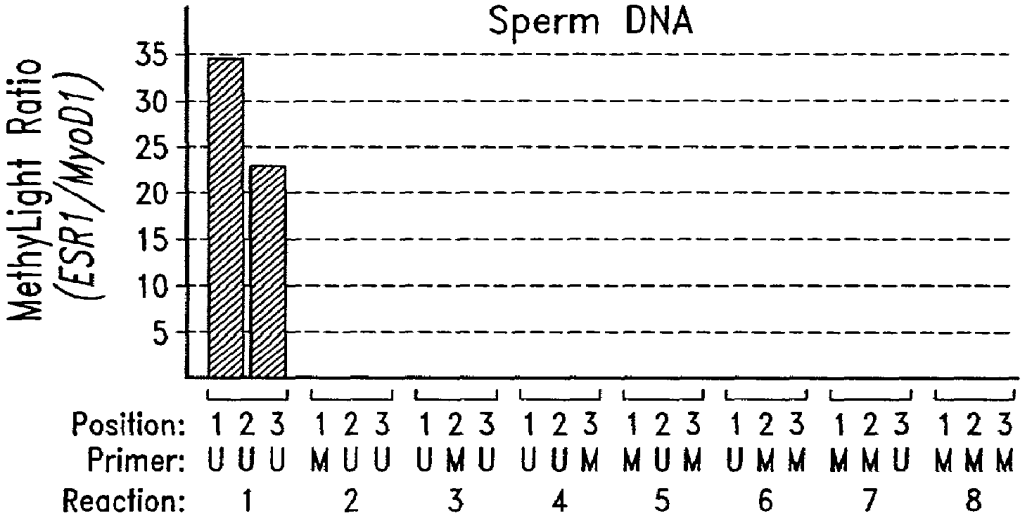


Fig. 6B

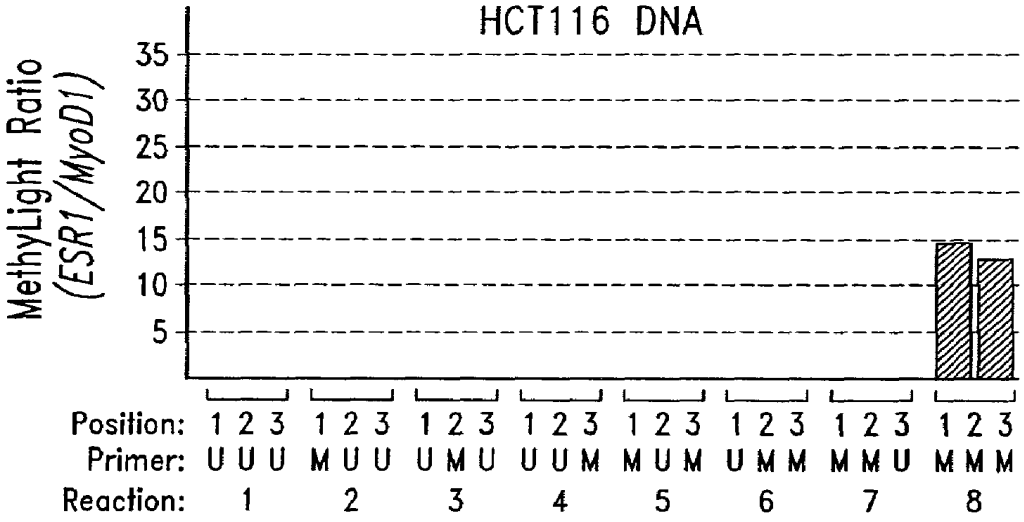


Fig. 6C

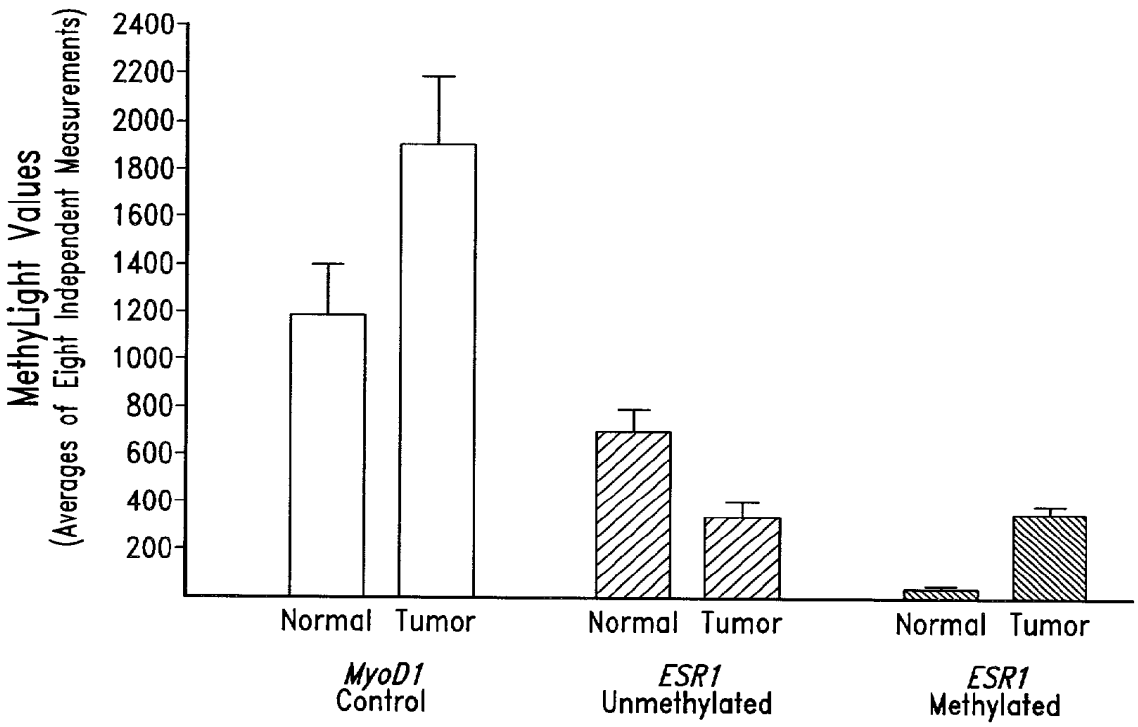

MethyLight Values
(Averages of Eight Independent Measurements)
2400
2200
2000
1800
1600
1400
1200
1000
800
600
400
200
Normal
Tumor
MyoD1
Control
Normal
Tumor
ESR1
Unmethylated
Normal
Tumor
ESR1
Methylated

Fig. 7

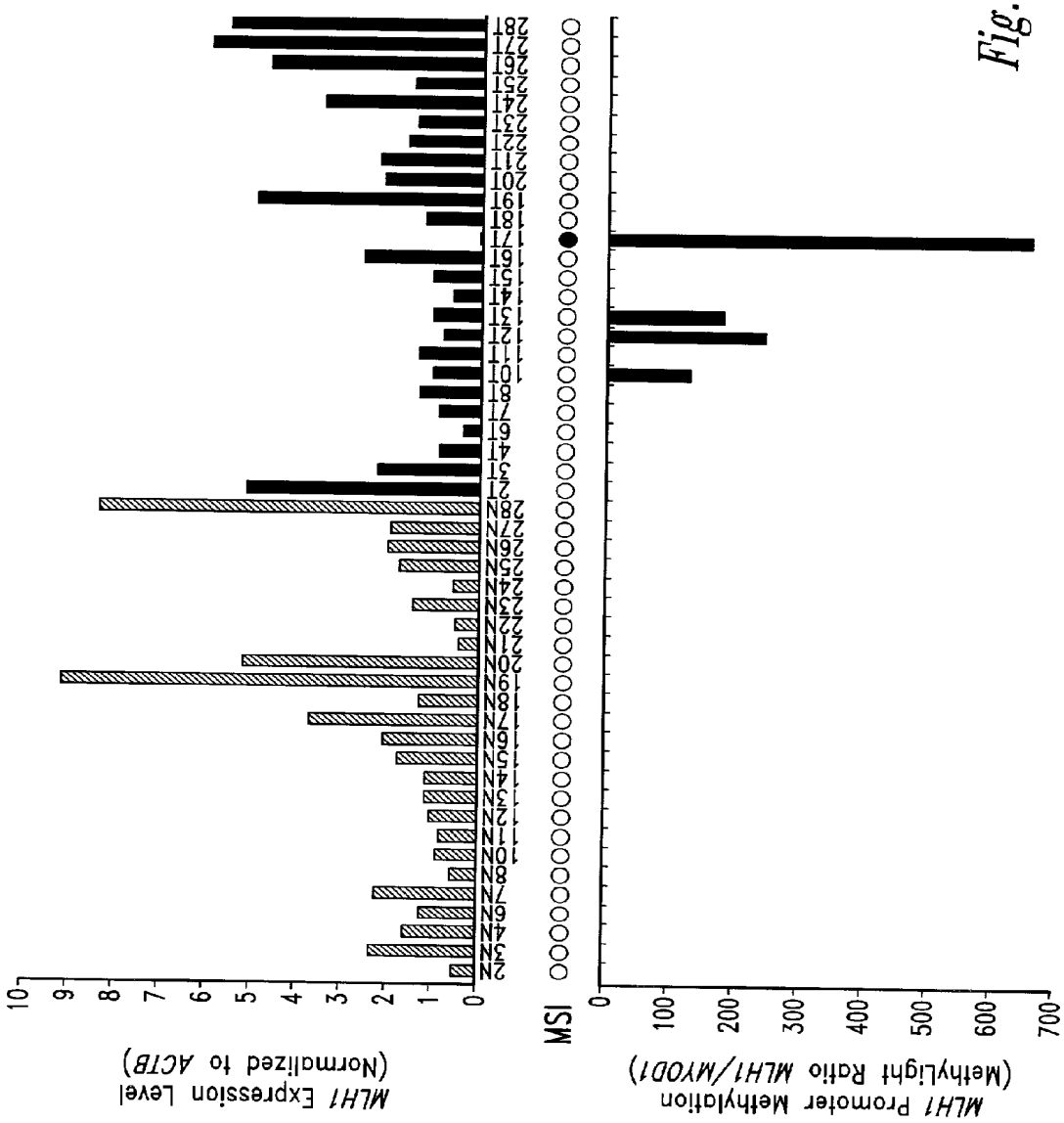

MLH1 Expression Level
(Normalized to ACTB)
MLH1 Promoter Methylation
(MethyLight Ratio MLH1/MYOD1)
MSI

Fig. 8

PROCESS FOR HIGH THROUGHPUT DNA METHYLATION ANALYSIS

TECHNICAL FIELD OF THE INVENTION

[0001] The present invention provides an improved high-throughput and quantitative process for determining methylation patterns in genomic DNA samples. Specifically, the inventive process provides for treating genomic DNA samples with sodium bisulfite to create methylation-dependent sequence differences, followed by detection with fluorescence-based quantitative PCR techniques.

BACKGROUND OF THE INVENTION

[0002] In higher order eukaryotic organisms, DNA is methylated only at cytosines located 5' to guanosine in the CpG dinucleotide. This modification has important regulatory effects on gene expression predominantly when it involves CpG rich areas (CpG islands) located in the promoter TN region of a gene sequence. Extensive methylation of CpG islands has been associated with transcriptional inactivation of selected imprinted genes and genes on the inactive X chromosome of females. Aberrant methylation of normally unmethylated CpG islands has been described as a frequent event in immortalized and transformed cells and has been frequently associated with transcriptional inactivation of tumor suppressor genes in human cancers.

[0003] DNA methylases transfer methyl groups from a universal methyl donor, such as S-adenosyl methionine, to specific sites on the DNA. One biological function of DNA methylation in bacteria is protection of the DNA from digestion by cognate restriction enzymes. Mammalian cellspossess methylases that methylate cytosine residues on DNA that are 5' neighbors of guanine CpG). This methylation may play a role in gene inactivation, cell differentiation, tumorigenesis, X-chromosome inactivation, and genomic imprinting. CpG islands remain unmethylated in normal cells, except during X-chromosome inactivation and parental specific imprinting where methylation of 5' regulatory regions can lead to transcriptional repression. DNA methylation is also a mechanism for changing the base sequence of DNA without altering its coding function. DNA methylation is a heritable, reversible and epigenetic change. Yet, DNA methylation has the potential to alter gene expression, which has profound developmental and genetic consequences.

[0004] The methylation reaction involves flipping a target cytosine out of an intact double helix to allow the transfer of a methyl group from S-adenosylmethionine in a cleft of the enzyme DNA (cystosine-5)-methyltransferase (Klimasauskas et al., *Cell* 76:357-369,-1994) to form 5-methylcytosine (5-mCyt). This enzymatic conversion is the only epigenetic modification of DNA known to exist in vertebrates and is essential for normal embryonic development (Bird, *Cell* 70:5-8, 1992; Laird and Jaenisch, *Human Mol. Genet.* 3:1487-1495, 1994; and Li et al., *Cell* 69:915-926, 1992). The presence of 5-mCyt at CpG dinucleotides has resulted in a 5-fold depletion of this sequence in the genome during vertebrate evolution, presumably due to spontaneous deamination of 5-mCyt to T (Schoreret et al., *Proc. Natl. Acad. Sci. USA* 89:957-961, 1992). Those areas of the genome that do not show such suppression are referred to as "CpG islands" (Bird, *Nature* 321:209-213, 1986; and Gardiner-Garden et al., *J. Mol. Biol* 196:261-282, 1987). These CpG island regions comprise about 1% of vertebrate genomes and also account for about 15% of the total number of CpG dinucleotides (Bird, *Nature* 321:209-213, 1986). CpG islands are typically between 0.2 to about 1 kb in length and are located upstream of many housekeeping and tissue-specific genes, but may also extend into gene coding regions. Therefore, it is the methylation of cytosine residues within CpG islands in somatic tissues, which is believed to affect gene function by altering transcription (Cedar, *Cell* 53:3-4, 1988).

[0005] Methylation of cytosine residues contained within CpG islands of certain genes has been inversely correlated with gene activity. This could lead to decreased gene expression by a variety of mechanisms including, for example, disruption of local chromatin structure, inhibition of transcription factor-DNA binding, or by recruitment of proteins which interact specifically with methylated sequences indirectly preventing transcription factor binding. In other words, there are several theories as to how methylation affects mRNA transcription and gene expression, but the exact mechanism of action is not well understood. Some studies have demonstrated an inverse correlation between methylation of CpG islands and gene expression, however, most CpG islands on autosomal genes remain unmethylated in the germline and methylation of these islands is usually independent of gene expression. Tissue-specific genes are usually unmethylated in the receptive target organs but are methylated in the germline and in non-expressing adult tissues. CpG islands of constitutively-expressed housekeeping genes are normally unmethylated in the germline and in somatic tissues.

[0006] Abnormal methylation of CpG islands associated-with tumor suppressor genes may also cause dcreased gene expression. Increased methylation of such regions may lead to progressive reduction of normal gene expression resulting in the selection of a population of cells having a selective growth advantage (i.e., a malignancy).

[0007] It is considered that an altered DNA methylation pattern, particularly methylation of cytosine residues, causes genome instability and is mutagenic. This, presumably, has led to an 80% suppression of a CpG methyl acceptor site in eukaryotic organisms, which methylate their genomes. Cytosine methylation further contributes to generation of polymorphism-and germ-line mutations and to transition mutations that inactivate tumor-suppressor genes (Jones, *Cancer Res.* 56:2463-2467, 1996). Methylation is also required for embryonic development of mammals (Li et al., *Cell* 69:915-926, 1992). It appears that the methylation of CpG-rich promoter regions may be blocking transcriptional activity. Ushijima et al. (*Proc. Natl. Acad. Sci. USA* 94:2284-2289,1997) characterized and cloned DNA fragments that show methylation changes during murine hepatocarcinogenesis. Data from a group of studies of altered methylation sites in cancer cells show that it is not simply the overall levels of DNA methylation that are altered in cancer, but changes in the distribution of methyl groups.

[0008] These studies suggest that methylation at CpG-rich sequences, known as CpG islands, provide an alternative pathway for the inactivation of tumor suppressors. Methylation of CpG oligonucleotides in the promoters of tumor suppressor genes can lead to their inactivation. Other studies provide data that alterations in the normal methylation process are associated with genomic instability (Lengauer et

al. *Proc. Natl. Acad. Sci. USA* 94:2545-2550, 1997). Such abnormal epigenetic changes may be found in many types of cancer and can serve as potential markers for oncogenic transformation, provided that there is a reliable means for rapidly determining such epigenetic changes. Therefore, there is a need in the art for a reliable and rapid (high-throughput) method for determining methylation as the preferred epigenetic alteration.

[0009] Methods to Determine DNA Methylation

[0010] There are a variety of genome scanning methods that have been used to identify altered methylation sites in cancer cells. For example, one method involves restriction landmark genomic scanning (Kawai et al., *Mol. Cell. Biol.* 14:7421-7427, 1994), and another example involves methylation-sensitive arbitrarily primed PCR (Gonzalgo et al., *Cancer Res.* 57:594-599, 1997). Changes in methylation patterns at specific CpG sites have been monitored by digestion of genomic DNA with methylation-sensitive restriction enzymes followed by Southern analysis of the regions of interest (digestion-Southern method). The digestion-Southern method is a straightforward method but it has inherent disadvantages in that it requires a large amount of high molecular weight DNA (at least or greater than 5 μg) and has a limited scope for analysis of CpG sites(as determined by the presence of recognition sites for methylation-sensitive restriction enzymes). Another method for analyzing changes in methylation patterns involves a PCR-based process that involves digestion of genomic DNA with methylation-sensitive restriction enzymes prior to PCR amplification (Singer-Sam et al., *Nucl. Acids Res.* 18:687, 1990). However, this method has not been shown effective because of a high degree of false positive signals (methylation present) due to inefficient enzyme digestion or over-amplification in a subsequent PCR reaction.

[0011] Genomic sequencing has been simplified for analysis of DNA methylation patterns and 5methylcytosine distribution by using bisulfite treatment (Frommer et al., *Proc. Natl. Acad. Sci. USA* 89:1827-1831, 1992). Bisulfite treatment of DNA distinguishes methylated from unmethylated cytosines; but original bisulfite genomic sequencing requires large-scale sequencing of multiple plasmid clones to determine overall methylation patterns, which prevents this technique from being commercially useful for determining methylation patterns in any type of a routine diagnostic assay.

[0012] In addition, other techniques have been reported which utilize bisulfite treatment of DNA as a starting point for methylation analysis. These include methylation-specific PCR (MSP) (Herman et al. *Proc. Natl. Acad. Sci. USA* 93:9821-9826, 1992); and restriction enzyme digestion of PCR products amplified from bisulfite-converted DNA (Sadri and Hornsby, *Nucl. Acids Res.* 24:5058-5059, 1996; and Xiong and Laird, *Nucl. Acids Res.* 25:2532-2534, 1997).

[0013] PCR techniques have been developed for detection of gene mutations (Kuppuswamy et al., *Proc. Natl. Acad. Sci. USA* 88:1143-1147, 1991) and quantitation of allelic-specific expression (Szabo and Mann, *Genes Dev.* 9:3097-3108, 1995; and Singer-Sam et al., *PCR Methods Appl.* 1:160-163, 1992). Such techniques use internal primers, which anneal to a PCR-generated template and terminate immediately 5' of the single nucleotide to be assayed. However an allelic-specific expression technique has not been tried within the context of assaying for DNA methylation patterns.

[0014] Most molecular biological techniques used to analyze specific loci, such as CpG islands in complex genomic DNA, involve some form of sequence-specific amplification, whether it is biological amplification by cloning in *E. coli*, direct amplification by PCR or signal amplification by hybridization with a probe that can be visualized. Since DNA methylation is added post-replicatively by a dedicated maintenance DNA methyltransferase that is not present in either *E. coli* or in the PCR reaction, such methylation information is lost during molecular cloning or PCR amplification. Moreover molecular hybridization does not discriminate between methylated and unmethylated DNA, since the methyl group on the cytosine does not participate in base pairing. The lack of a facile way to amplify the methylation information in complex genomic DNA has probably been a most important impediment to DNA methylation research. Therefore, there is a need in the art to improve upon methylation detection techniques, especially in a quantitative manner.

[0015] The indirect methods for DNA methylation pattern determinations at specific loci that have been developed rely on techniques that alter the genomic DNA in a methylation-dependent manner before the amplification event. There- are two primary methods that have been utilized to achieve this methylation-dependent DNA alteration. The first is digestion by a restriction enzyme that is affected in its activity by 5-methylcytosine in a CpG sequence context. The cleavage, or lack of it, can subsequently be revealed by Southern blotting or by PCR. The other technique that has received recent widespread use-is the treatment of genomic DNA with sodium bisulfite. Sodium bisulfite treatment converts all unmethylated cytosines in the DNA to uracil by deamination, but leaves the methylated cytosine residues intact. Subsequent PCR amplification replaces the uracil residues with thymines and the 5-methylcytosine residues with cytosines. The resulting sequence difference has been detected using standard DNA sequence detection techniques, primarily PCR.

[0016] Many DNA methylation detection techniques utilize bisulfite treatment. Currently, all bisulfite treatment-based methods are followed by a PCR reaction to analyze specific loci within the genome. There are two principally different ways in which the sequence difference generated by the sodium bisulfite treatment can be revealed. The first is to design PCR primers that uniquely anneal with either methylated or unmethylated converted DNA. This technique is referred to as "methylation specific PCR" or "MSP". The method used by all other bisulfite-based techniques (such as bisulfite genomic sequencing, COBRA and Ms-SNuPE) is to amplify the bisulfite-converted DNA using primers that anneal at locations that lack CpG dinucleotides in the original genomic sequence. In this way, the PCR primers can amplify the sequence in between the two primers, regardless of the DNA methylation status of that sequence in the original genomic DNA. This results in a pool of different PCR products, all with the same length and differing in their sequence only at the sites of potential DNA methylation at CpGs located in between the two primers. The difference between these methods of processing the bisulfite-converted sequence is that in MSP, the methylation information is

derived from the occurrence or lack of occurrence of a PCR product, whereas in the other techniques a mix of products is always generated and the mixture is subsequently analyzed to yield quantitative information on the relative occurrence of the different methylation states.

[0017] MSP is a qualitative technique. There are two reasons that it is not quantitative. The first is that methylation information is derived from the comparison of two separate PCR reactions (the methylated and the unmethylated version). There are inherent difficulties in making kinetic comparisons of two different PCR reactions. The other problem with MSP is that often the primers cover more than one CpG dinucleotide. The consequence is that multiple sequence variants can be generated, depending on the DNA methylation pattern in the original genomic DNA. For instance, if the forward primer is a 24-mer oligonucleotide that covers 3 CpGs, then 2^3=8 different theoretical sequence permutations could arise in the genomic DNA following bisulfite conversion within this 24-nucleotide sequence. If only a fully methylated and a fully unmethylated reaction is run, then you are really only investigating 2 out of the 8 possible methylation states. The situation is fisher complicated if the intermediate methylation states lead to amplification, but with reduced efficiency. Therefore, the MSP technique is non-quantitative. Therefore, there is a need in the art to improve the MSP technique and change it to be more quantitaive and facilitate its process to greater throughput. The present invention addresses this need or a more rapid and quantitative methylation assay.

SUMMARY OF THE INVENTION

[0018] The present invention provides a method for detecting a methylated CpG island within a genomic sample of DNA comprising:

[0019] (a) contacting a genomic sample of DNA from a patient with a modifying agent that modifies unmethylated cytosine to produce a converted nucleic acid;

[0020] (b) amplifying the converted nucleic acid by means of two oligonucleotide primers in the presence or absence of one or a plurality of specific oligonucleotide probes, wherein one or more of the oligonucleotide primers and/or probes are capable of distinguishing between unmethylated and methylated nucleic acid; and

[0021] (c) detecting the methylated nucleic acid based on amplification-mediated displacement of the probe. Preferably, the amplifying step is a polymerase chain reaction (PCR) and the modifying agent is bisulfite. Preferably, the converted nucleic acid contains uracil in place of unmethylated cytosine residues present in the unmodified genomic sample of DNA. Preferably, the probe further comprises a fluorescence label moiety and the amplification and detection step comprises fluorescence-based quantitative PCR.

[0022] The invention provides a method for detecting a methylated CpG-containing nucleic acid comprising:

[0023] (a) contacting a nucleic acid-containing sample with a modifying agent that modifies unmethylated cytosine to produce a converted nucleic acid;

[0024] (b) amplifying the converted nucleic acid in the sample by means of oligonucleotide primers in the presence of a CpG-specific oligonucleotide probe, wherein the CpG-specific probe, but not the primers, distinguish between modified unmethylated and methylated nucleic acid; and

[0025] (c) detecting the methylated nucleic acid based upon an amplification-mediated displacement of the CpG-specific probe. Preferably, the amplifying step comprises a polymerase chain reaction (PCR) and the modifying agent comprises bisulfite. Preferably, the converted nucleic acid contains uracil in place of unmethylated cytosine residues present in the unmodified nucleic acid-containing sample. Preferably, the detection method is by means of a measurement of a fluorescence signal based on amplification-mediated displacement of the CpG-specific probe and the amplification and detection method comprises fluorescence-based quantitative PCR. The methylation amounts in the nucleic acid sample are quantitatively determined based on reference to a control reaction for amount of input nucleic acid.

[0026] The present invention further provides a method for detecting a methylated CpG-containing nucleic acid comprising:

[0027] (a) contacting a nucleic acid-containing sample with a modifying agent that modifies unmethylated cytosine to produce a converted nucleic acid;

[0028] (b) amplifying the converted nucleic acid in the sample by means of oligonucleotide primers and in the presence of a CpG-specific oligonucleotide probe, wherein both the primers and the CpG-specific probe distinguish between modified unmethylated and methylated nucleic acid; and

[0029] (c) detecting the methylated nucleic acid based on amplification-mediated displacement of the CpG-specific probe. Preferably, the amplifying step is a polymerase chain reaction (PCR) and the modifying agent is bisulfite. Preferably, the converted nucleic acid contains uracil in place of unmethylated cytosine residues present in the unmodified nucleic acid-containing sample. Preferably, the detection method comprises measurement of a fluorescence signal based on amplification-mediated displacement of the CpG-specific probe and the amplification and detection method comprises fluorescence-based quantitative PCR.

[0030] The present invention further provides a methylation detection kit useful for the detection of a methylated CpG-containing nucleic acid comprising a carrier means being compartmentalized to receive in close confinement therein one or more containers comprising:

[0031] (i) a first container containing a modifying agent that modifies unmethylated cytosine to produce a converted nucleic acid;

[0032] (ii) a second container containing primers for amplification of the converted nucleic acid;

[0033] (iii) a third container containing primers for the amplification of control unmodified nucleic acid; and

[0034] (iv) a fourth container containing a specific oligonucleotide probe the detection of which is based on amplification-mediated displacement,

[0035] wherein the primers and probe each may or may not distinguish between unmethylated and methylated nucleic acid. Preferably, the modifying agent comprises bisulfite. Preferably, the modifying agent converts cytosine residues to uracil residues. Preferably, the specific oligonucleotide probe is a CpG-specific oligonucleotide probe, wherein the probe, but not the primers for amplification of the converted nucleic acid, distinguishes between modified unmethylated and methylated nucleic acid. Alternatively, the specific oligonucleotide probe is a CpG-specific oligonucleotide probe, wherein both the probe and the primers for amplification of the converted nucleic acid, distinguish between modified unmethylated and methylated nucleic acid. Preferably, the probe further comprises a fluorescent moiety linked to an oligonucleotide base directly or through a linker moiety and the probe is a specific, dual-labeled TaqMan probe.

BRIEF DESCRIPTION OF THE DRAWINGS

[0036] **FIG. 1** shows an outline of the MSP technology (prior art) using PCR primers that initially discriminate between methylated and unmethylated (bisulfite-converted) DNA. The top part shows the result of the MSP process when unmethylated single-stranded genomic DNA is initially subjected to sodium bisulfite conversion (deamination of unmethylated cytosine residues to uracil) followed by PCR reactions with the converted template, such that a PCR product appears only with primers specifically annealing to converted (and hence unmethylated) DNA. The bottom portion shows the contrasting result when a methylated single-stranded genomic DNA sample is used. Again, the process first provides for bisulfite treatment followed by PCR reactions such that a PCR product appears only with primers specifically annealing to unconverted (and hence initially methylated) DNA.

[0037] **FIG. 2** shows an alternate process for evaluating DNA methylation with sodium bisulfite-treated genomic DNA using nondiscriminating (with respect to methylation status) forward and reverse PCR primers to amplify a specific locus. In this illustration, denatured (ie., single-stranded) genomic DNA is provided that has mixed methylation status, as would typically be found in a sample for analysis. The sample is converted in a standard sodium bisulfite reaction and the mixed products are amplified by a PCR reaction using primers that do not overlap any CpG dinucleotides. This produces an unbiased (with respect to methylation status) heterogeneous pool of PCR products. The mixed or heterogeneous pool can then be analyzed by a technique capable of detecting sequence differences, including direct DNA sequencing, subcloning of PCR fragments followed by sequencing of representative clones, single-nucleotide primer extension reaction (MS-SNuPE), or restriction enzyme digestion (COBRA).

[0038] **FIG. 3** shows a flow diagram of the inventive process in several, but not all, alternative embodiments for PCR product analysis. Variations in detection methodology, such as the use of dual probe technology (Lightcycler®) or flourescent primers (Sunrise® technology) are not shown in this Figure. Specifically, the inventive process begins with a mixed sample of genomic DNA that is converted in a sodium bisulfite reaction to a mixed pool of methylation-dependent sequence differences according to standard procedures (the bisulfite process converts unmethylated cytosine residues to uracil). Fluorescence-based PCR is then performed either in an "unbiased" PCR reaction with primers that do not overlap known CpG methylation sites (left arm of **FIG. 3**), or in a "biased" reaction with PCR primers that overlap known CpG dinucleotides (right arm of **FIG. 3**). Sequence discrimination can occur either at the level of the amplification process (C and D) or at the level of the fluorescence detection process (B), or both (D). A quantitative test for methylation patterns in the genomic DNA sample is shown on the left arm (B), wherein sequence discrimination occurs at the level of probe hybridization. In this version, the PCR reaction provides for unbiased amplification in the presence of a fluorescent probe that overlaps a particular putative methylation site. An unbiased control for the amount of input DNA is provided by a reaction in which neither the primers, nor the probe overlie, any CpG dinucleotides (A). Alternatively, as shown in the right arm of **FIG. 3, a** qualitative test for genomic methylation is achieved by probing of the biased PCR pool with either control, oligonucleotides that do not "cover" known methylation sites (C; a fluorescence-based version of the MSP technique), or with oligonucleotides covering potential methylation sites (D).

[0039] **FIG. 4** shows a flow chart overview of the inventive process employing a "TaqMan®" probe in the amplification process. Briefly, double-stranded genomic DNA is treated with sodium bisulfite and subjected to one of two sets of PCR reactions using TaqMan® probes; namely with either biased primers and TaqMan® probe (left column), or unbiased primers and TaqMan® probe (right column). The TaqMan® probe is dual-labeled with a fluorescent "reporter" (labeled "R" in **FIG. 4**) and "qencher" (labeled "O") molecules, and is designed to be specific for a relatively high GC content region so that it melts out at about 10° C. higher temperature in the PCR cycle than the forward or reverse primers. This allows it to remain fully hybridized during the PCR annealing/extension step. As the Taq polymerase enzymatically synthesizes a new strand during PCR, it will eventually reach the annealed TaqMan® probe. The Taq polymerase 5' to 3' endonuclease activity will then displace the TaqMan® probe by digesting it to release the fluorescent reporter molecule for quantitative detection of its now unquenched signal using a real-time fluorescent system as described herein.

[0040] **FIG. 5** shows a comparison of the inventive assay to a conventional COBRA assay. Panel A shows a COBRA gel used to determine the level of DNA methylation at the ESR1 locus in DNAs of known methylation status (sperm, unmethylated) and HCT116 (methylated). The relative amounts of the cleaved products are indicated below the gel. A 56-bp fragment represents DNA molecules in which the TaqI site proximal to the hybridization probe is methylated in the original genomic DNA. The 86-bp fragment represents DNA molecules in which the proximal TaqI site is unmethylated and the distal site is methylated. Panel B summarizes the COBRA results and compares them to results obtained with the methylated and unmethylated version of the inventive assay process. The results are expressed as ratios between the methylation-specific reactions and a control reaction. For the bisulfite-treated samples, the control reaction was a MYOD1 assay as described in Example

1. For the untreated samples, the ACTB primers described for the RT-PCR reactions were used as a control to verify the input of unconverted DNA samples. (The ACTB primers do not span an intron). "No PCR" indicates that no PCR-product was obtained on unconverted genomic DNA with COBRA primers designed amplify bisulfite-converted DNA sequences.

[0041] **FIG. 6** illustrates a determination of the specificity of the oligonucleotides. Eight different combinations of forward primer, probe and reverse primer were-tested on DNA samples with known methylation or lack of methylation at the ESR1 locus. Panel A shows the nomenclature used for the combinations of the ESR1 oligos. "U" refers to the oligo sequence that anneals with bisulfite-converted unmethylated DNA, while "M" refers to the methylated version. Position 1 indicates the forward PCR primer, position 2 the probe, and position 3 the reverse primer. The combinations used for the eight reactions are shown below each pair of bars, representing duplicate experiments. The results are expressed as ratios between the and the MYOD1 control values. Panel B represents an analysis of human sperm DNA. Panel C represents an analysis of DNA obtained from the human colorectal cancer cell line HCT116.

[0042] **FIG. 7** shows a test of the reproducibility of the reactions. Assays were performed in eight independent reactions to determine the reproducibility on samples of complex origin. A primary human colorectal adenocarcinoma and matched normal mucosa was used for this purpose (samples 10N and 10T shown in **FIG. 8**). The results shown in this figure represent the raw values obtained in the assay. The values have been plate-normalized, but not corrected for input DNA. The bars indicate the mean values obtained for the eight separate reactions. The error bars represent the standard error of the mean.

[0043] **FIG. 8** illustrates a comparison of MLH1 expression, microsatellite instability and MLH1 promoter methylation of 25 matched-paired human colorectal samples. The upper chart shows the MLH1 expression levels measured by quantitative, real time RT-PCR (TaqMan®) in matched normal (hatched bars) and tumor (solid black bars) colorectal samples. The expression levels are displayed as a ratio between MLH1 and ACTB measurements. Microsatellite instability status (MSI) is indicated by the circles located between the two charts. A black circle denotes MSI positivity, while an open circle indicates that the sample is MSI negative, as determined by analysis of the BAT25 and BAT26 loci. The lower chart shows the methylation status of the MLH1 locus as determined by an inventive process. The methylation levels are represented as the ratio between the MLH1 methylated reaction and the MYOD1 reaction.

DETAILED DESCRIPTION OF THE INVENTION

[0044] The present invention provides a rapid, sensitive, reproducible high-throughput method for detecting methylation patterns in samples of nucleic acid. The invention provides for methylation-dependent modification of the nucleic acid, and then uses processes-of nucleic acid amplification, detection, or both to distinguish between methylated and unmethylated residues present in the original sample of nucleic acid. In a preferred embodiment, the invention provides for determining the methylation status of CpG islands within samples of genomic DNA.

[0045] In contrast to previous methods for determining methylation patterns, detection of the methylated nucleic acid is relatively rapid and is based on amplification-mediated displacement of specific oligonucleotide probes. In a preferred embodiment, amplification and detection, in fact, occur simultaneously as measured by fluorescence-based real-time quantitative PCR ("RT-PCR") using specific, dual-labeled TaqMan® oligonucleotide probes. The displaceable probes can be specifically designed to distinguish between methylated and unmethylated CpG sites present in the original, unmodified nucleic acid sample.

[0046] Like the technique of methylation-specific PCR ("MSP"; U.S. Pat. No. 5,786,146), the present invention provides for significant advantages over previous PCR-based and other methods (e.g., Southern analyses) used for determining methylation patterns. The present invention is substantially more sensitive than Southern analysis, and facilitates the detection of a low number (percentage) of methylated alleles in very small nucleic acid samples, as well as paraffin-embedded samples. Moreover, in the case of genomic DNA, analysis is not limited to DNA sequences recognized by methylation-sensitive restriction endonucleases, thus allowing for fine mapping of methylation patterns across broader CpG-rich regions. The present invention also eliminates the any false-positive results, due to incomplete digestion by methylation-sensitive restriction enzymes, inherent in previous PCR-based methylation methods.

[0047] The present invention also offers significant advantages over MSP technology. It can be applied as a quantitative process for measuring methylation amounts, and is substantially more rapid. One important advance over MSP technology is that the gel electrophoresis is not only a time-consuming manual task that limits high throughput capabilities, but the manipulation and opening of the PCR reaction tubes increases the chance of sample mis-identification and it greatly increases the chance of contaminating future PCR reactions with trace PCR products. The standard method of avoiding PCR contamination by uracil incorporation and the use of Uracil DNA Glycosylase (AmpErase) is incompatible with bisulfite technology, due to the presence of uracil in bisulfite-treated DNA. Therefore, the avoidance of PCR product contamination in a high-throughput application with bisulfite-treated DNA is a greater technical challenge than for the amplification of unmodified DNA. The present invention does not require any post-PCR manipulation or processing. This not only greatly reduces the amount of labor involved in the analysis of bisulfite-treated DNA, but it also provides a means to avoid handling of PCR products that could contaminate future reactions.

[0048] Two factors limit MSP to, at best, semi-quantitative applications. First, MSP methylation information is derived from the comparison of two separate PCR reactions (the methylated and the unmethylated versions). There are inherent difficulties in making kinetic comparisons of two different PCR reactions without a highly quantitative method of following the amplification reaction, such as Real-Time Quantitative PCR. The other problem relates to the fact that MSP amplification is provided for by, means of particular CpG-specific oligonucleotides; that is, by biased primers.

Often, the DNA sequence covered by such primers contains more than one CpG dinucleotide with the consequence that the sequence amplified will represent only one of multiple potential sequence variants present, depending on the DNA methylation pattern in the original genomic DNA. For instance, if the forward primer is a 24-mer oligonucleotide that covers 3 CpGs, then 2^3=8 different theoretical sequence permutations could arise in the genomic DNA following bisufite conversion within this 24-nucleotide sequence. If only a fully methylated and a fully unmethylated reaction is run, then only 2 out of the 8 possible methylation states are analyzed.

[0049] The situation is further complicated if the intermediate methylation states are non-specifically amplified by the fully methylated or fully unmethylated primers. Accordingly, the MSP patent explicitly describes a non-quantitative technique based on the occurrence or non-occurrence of a PCR product in the fully methylated , versus fully unmethylated reaction, rather than a comparison of the kinetics of the two reactions.

[0050] By contrast, one embodiment of the present invention provides for the unbiased amplification of all possible methylation states using primers that do not cover any CpG sequences in the original, unmodified DNA sequence. To the extent that all methylation patterns are amplified equally, quantitative information about DNA methylation patterns can then be distilled from the resulting PCR pool by any technique capable of detecting sequence differences (e.g., by fluorescence-based PCR).

[0051] Furthermore, the present invention is substantially faster than MSP. As indicated above, MSP relies on the occurrence or non-occurrence of a PCR product in the methylated, versus unmethylated reaction to determine the methylation status of a CpG sequence covered by a primer. Minimally, this requires performing agarose or polyacrylamide gel electrophoretic analysis (see U.S. Pat. No. 5,786, 146, FIGS. **2**A-**2**E, and **3**A-**3**E). Moreover, determining the methylation status of any CpG sites within a given MSP amplified region would require additional analyses such as: (a) restriction endonuclease analysis either before, or after (e.g., COBRA analysis; Xiong and Laird, *Nucleic Acids Res.* 25:2532-2534, 1997) nucleic acid modification and amplification, provided that either the unmodified sequence region of interest contains methylation-sensitive sites, or that modification (e.g., bisulfite) results in creating or destroying restriction sites; (b) single nucleotide primer extension reactions (Ms-SNuPE; Gonzalo and Jones, *Nucleic Acids Res* 25: 2529-2531, 1997); or (c) DNA sequencing of the amplification products. Such additional analyses are not only subject to error (incomplete restriction enzyme digestion), but also add substantial time and expense to the process of determining the CpG methylation status of, for example, samples of genomic DNA.

[0052] By contrast, in a preferred embodiment of the present invention, amplification and detection occur simultaneously as measured by fluorescence-based real-time quantitative PCR using specific, dual-labeled oligonucleotide probes. In principle, the methylation status at any probe-specific sequence within an amplified region can be determined contemporaneously with amplification, with no requirement for subsequent manipulation or analysis.

[0053] As disclosed by MSP inventors, "[t]he only technique that can provide more direct analysis than MSP for most CpG sites within a defined region is genomic sequencing." (U.S. Pat. No. 5,786,146 at 5,line 15-17). The present invention provides, in fact, a method for the partial direct sequencing of modified CpG sites within a known (previously sequenced) region of genomic DNA. Thus, a series of CpG-specific TaqMan® probes, each corresponding to a particular methylation site in a given amplified DNA region, are constructed. This series of probes are then utilized in parallel amplification reactions, using aliquots of a single, modified DNA sample, to simultaneously determine the complete methylation pattern present in the original unmodified sample of genomic DNA. This is accomplished in a fraction of the time and expense required for direct sequencing of the sample of genomic DNA, and are substantially more sensitive. Moreover, one embodiment of the present invention provides for a quantitative assessment of such a methylation pattern.

[0054] The present invention has identified four process techniques and associated diagnostic kits, utilizing a methylation-dependent nucleic acid modifying agent (e.g., bisulfite), to both qualitatively and quantitatively determine CpG methylation status in nucleic acid samples (e.g., genomic DNA samples). The four processes are outlined in **FIG. 3** and labeled at the bottom with the letters A through D. Overall, methylated-CpG sequence discrimination is designed to occur at the level of amplification, probe hybridization or at both levels. For example, applications C and D utilize "biased" primers that distinguish between modified unmethylated and methylated nucleic acid and provide methylated-CpG sequence discrimination at the PCR amplification level. Process B uses "unbiased" primers (that do not cover CpG methylation sites), to provide for unbiased amplification of modified nucleic acid, but rather utilize probes that distinguish between modified unmethylated and methylated nucleic acid to provide for quantitative methylated-CpG sequence discrimination at the detection level (e.g., at the fluorescent (or luminescent) probe hybridization level only). Process A does not, in itself, provide for methylated-CpG sequence discrimination at either the amplification or detection levels, but supports and validates the other three applications by providing control reactions for input DNA.

[0055] Process D.

[0056] In a first embodiment (**FIG. 3**, Application D), the invention provides a method for qualitatively detecting a methylated CpG-containing nucleic acid, the method including: contacting a nucleic acid-containing sample with a modifying agent that modifies unmethylated cytosine to produce a converted nucleic acid; amplifying the converted nucleic acid by means of two oligonucleotide primers in the presence of a specific oligonucleotide hybridization probe, wherein both the primers and probe distinguish between modified, unmethylated and methylated nucleic acid; and detecting the "methylated" nucleic acid based on amplification-mediated probe displacement.

[0057] The term "modifies" as used herein means the conversion of an unmethylated cytosine to another nucleotide by the modifying agent, said conversion distinguishing unmethylated from methylated cytosine in the original nucleic acid sample. Preferably, the agent modifies unmethylated cytosine to uracil. Preferably, the agent used for modifying unmethylated cytosine is sodium bisulfite, how-

ever, other equivalent modifying agents that selectively modify unmethylated cytosine, but not methylated cytosine, can be substituted in the method of the invention. Sodium-bisulfite readily reacts with the 5,6-double bond of cytosine, but not with methylated cytosine, to produce a sulfonated cytosine intermediate that undergoes deamination under alkaline conditions to produce uracil (Example 1). Because Taq polymerase recognizes uracil as thymine and 5-methylcytidine (m5C) as cytidine, the sequential combination of sodium bisulfite treatment and PCR amplification results in the ultimate conversion of unmethylated cytosine residues to thymine (C→U→T) and methylated cytosine residues ("mC") to cytosine (mC→mC→C). Thus, sodium-bisulfite treatment of genomic DNA creates methylation-dependent sequence differences by converting unmethylated cyotsines to uracil, and upon PCR the resultant product contains cytosine only at positions where methylated cytosine occurs in the unmodified nucleic acid.

[0058] Oligonucleotide "primers," as used herein, means linear, single-stranded, oligomeric deoxyribonucleic or ribonucleic acid molecules capable of sequence-specific hybridization (annealing) with complementary strands of modified or unmodified nucleic acid. As used herein, the specific primers are preferably DNA. The primers of the invention embrace oligonucleotides of appropriate sequence and sufficient length so as to provide for specific and efficient initiation of polymerization (primer extension) during the amplification process. As used in the inventive processes, oligonucleotide primers typically contain 12-30 nucleotides or more, although may contain fewer nucleotides. Preferably, the primers contain from 18-30 nucleotides. The exact length will depend on multiple factors including temperature (during amplification), buffer, and nucleotide composition. Preferably, primers are single-stranded although double-stranded primers may be used if the strands are first separated. Primers may be prepared using any suitable method, such as conventional phosphotriester and phosphodiester methods or automated embodiments which are commonly known in the art.

[0059] As used in the inventive embodiments herein, the specific primers are preferably designed to be substantially complementary to each strand of the genomic locus of interest. Typically, one primer is complementary to the negative, (−) strand of the locus (the "lower" strand of a horizontally situated double-stranded DNA molecule) and the other is complementary to the positive (+) strand ("upper" strand). As used in the embodiment of Application D, the primers are preferably designed to overlap potential sites of DNA methylation (CpG nucleotides) and specifically distinguish modified unmethylated from methylated DNA. Preferably, this sequence discrimination is based upon the differential annealing temperatures of perfectly matched, versus mismatched oligonucleotides. In the embodiment of Application D, primers are typically designed to overlap from one to several CpG sequences. Preferably, they are designed to overlap from 1 to 5 CpG sequences, and most preferably from 1 to 4 CpG sequences. By contrast, in a quantitative embodiment of the invention, the primers do not overlap any CpG sequences.

[0060] In the case of fully "unmethylated" (complementary to modified unmethylated nucleic acid strands) primer sets, the anti-sense primers contain adenosine residues ("As") in place of guanosine residues ("Gs") in the corresponding (−) strand sequence. These substituted As in the anti-sense primer will be complementary to the uracil and thymidine residues ("Us" and "Ts") in the corresponding (+) strand region resulting from bisulfite modification of unmethylated C residues ("Cs") and subsequent amplification. The sense primers, in this case, are preferably designed to be complementary to anti-sense primer extension products, and contain Ts in place of unmethylated Cs in the corresponding (+) strand sequence. These substituted Ts in the sense primer will be complementary to the As, incorporated in the anti-sense primer extension products at positions complementary to modified Cs (Us) in the original (+) strand.

[0061] In the case of fully-methylated primers (complementary to methylated CpG-containing nucleic acid strands), the anti-sense primers will not contain As in place of Gs in the corresponding (−) strand sequence that are complementary to methylated Cs (i.e., mCpG sequences) in the original (+) strand. Similarly, the sense primers in this case will not contain Ts in place of methylated Cs in the corresponding (+) strand mCpG sequences. However, Cs that are UP not in CpG sequences in regions covered by the fully-methylated primers, and are not methylated, will be represented in the fully-methylated primer set as described above for unmethylated primers.

[0062] Preferably, as employed in the embodiment of Application D, the amplification process provides for amplifying bisulfite converted nucleic acid by means of two oligonucleotide primers in the presence of a specific oligonucleotide hybridization probe. Both the primers and probe distinguish between modified unmethylated and methylated nucleic acid. Moreover, detecting the "methylated" nucleic acid is based upon amplification-mediated probe fluorescence. In one embodiment, the fluorescence is generated by probe degradation by 5' to 3' exonuclease activity of the polymerase enzyme. In another embodiment, the fluorescence is generated by fluorescence energy transfer effects between two adjacent hybridizing probes (Lightcycler® technology) or between a hybridizing probe and a primer. In another embodiment, the fluorescence is generated by the primer itself (Sunrise® technology). Preferably, the amplification process is an enzymatic chain reaction that uses the oligonucleotide primers to produce exponential quantities of amplification-product, from a target locus, relative to the number of reaction steps involved.

[0063] As describe above, one member of a primer set is complementary to the (−) strand, while the other is complementary to the (+) strand. The primers are chosen to bracket the area of interest to be amplified; that is, the "amplicon." Hybridization of the primers to denatured target nucleic acid followed by primer extension with a DNA polymerase and nucleotides, results in synthesis of new nucleic acid strands corresponding to the amplicon. Preferably, the DNA polymerase is Taq polymerase, as commonly used in the art. Although equivalent polymerases with a 5' to 3' nuclease activity can be substituted. Because the new amplicon sequences are also templates for the primers and polymerase, repeated cycles of denaturing, primer annealing, and extension results in exponential production of the amplicon. The product of the chain reaction is a discrete nucleic acid duplex, corresponding to the amplicon sequence, with termini defined by the ends of the specific primers employed. Preferably the amplification method used is that of PCR (Mullis et al., *Cold Spring Harb. Symp. Quant. Biol.* 51:263-

273; Gibbs, *Anal. Chem.* 62:1202-1214, 1990), or more preferably, automated embodiments thereof which are commonly known in the art.

[**0064**] Preferably, methylation-dependent sequence differences are detected by methods based on fluorescence-based quantitative PCR (real-time quantitative PCR, Heid et al., *Genome Res.* 6:986-994, 1996; Gibson et al., *Genome Res.* 6:995-1001, 1996) (e.g., "TaqMan®,""Lightcycler®," and "Sunrise®" technologies). For the TaqMan® and Lightcycler® technologies, the sequence discrimination can occur at either or both of two steps: (1) the amplification step, or (2) the fluorescence detection step. In the case of the "Sunrise®" technology, the amplification and fluorescent steps are the same. In the case of the FRET hybridization, probes format on the Lightcycler®, either or both of the FRET oligonucleotides can be used to distinguish the sequence difference. Most preferably the amplification process, as employed in all inventive embodiments herein, is that of fluorescence-based Real Time Quantitative PCR (Heid et al., *Genome Res.* 6:986-994, 1996) employing a dual-labeled fluorescent oligonucleotide probe (TaqMan® PCR, using an ABI Prism 7700 Sequence Detection System, Perkin Elmer Applied Biosystems, Foster City, Calif.).

[**0065**] The "TaqMan®" PCR reaction uses a pair of amplification primers along with a nonextendible interrogating oligonucleotide, called a TaqMan® probe, that is designed to a hybridize to a GC-rich sequence located between the forward and reverse (ie., sense and anti-sense) primers. The TaqMan® probe further comprises a fluorescent "reporter moiety" and a "quencher moiety" covalently bound to linker moieties (e.g., phosphoramidites) attached to nucleotides of the TaqMan® oligonucleotide. Examples of suitable reporter and quencher molecules are: the 5' fluorescent reporter dyes 6FAM ("FAM"; 2,7 dimeihoxy-4,5-dichloro-6-carboxy-fluorescein), and TET (6-carboxy4,7,2',7'-tetrachlorofluorescein); and the 3' quencher dye TAMRA (6-carboxytetramethylrhodamine) (Livak et al., *PCR Methods Appl.* 4:357-362, 1995; Gibson et al., *Genome Res.* 6:995-1001; and 1996; Heid et al., *Genome Res.* 6:986-994, 1996).

[**0066**] One process for designing appropriate TaqMan®) probes involves utilizing a software facilitating tool, such as "Primer Express" that can determine the variables of CpG island location within GC-rich sequences to provide for at least a 10° C. melting temperature difference (relative to the primer melting temperatures) due to either specific sequence (tighter bonding of GC, relative to AT base pairs), or to primer length.

[**0067**] The TaqMan® probe may or may not cover known CpG methylation sites, depending on the particular inventive process used. Preferably, in the embodiment of Application D, the TaqMan® probe is designed to distinguish between modified unmethylated and methylated nucleic acid by overlapping from 1 to 5 CpG sequences. As described above for the fully unmethylated and fully methylated primer sets, TaqMan® probes may be designed to be complementary to either unmodified nucleic acid, or, by appropriate base substitutions, to bisulfite-modified sequences that were either fully unmethylated or fully methylated in the original, unmodified nucleic acid sample.

[**0068**] Each oligonucleotide primer or probe in the TaqMan® PCR reaction can span anywhere from zero to many different CpG dinucleotides that each can result in two different sequence variations following bisulfite treatment (^{m}CpG, or UpG). For instance, if an oligonucleotide spans 3 CpG dinucleotides, then the number of possible sequence variants arising in the genomic DNA is 2^3=8 different sequences. If the forward and reverse primer each span 3 CpGs and the probe oligonucleotide (or both oligonucleotides together in the case of the FRET format) spans another 3, then the total number of sequence permutations becomes 8×8×8=512. In theory, one could design separate PCR reactions to quantitatively analyze the relative amounts of each of these 512 sequence variants. In practice, a substantial amount of qualitative methylation information can be derived from the analysis of a much smaller number of sequence variants. Thus, in its most simple form, the inventive process can be performed by designing reactions for the fully methylated and the fully unmethylated variants that represent the most extreme sequence variants in a hypothetical example (see **FIG. 3**, Application D). The ratio between these two reactions, or alternatively the ratio between the methylated reaction and a control reaction **(FIG. 3**, Application A), would provide a measure for the level of DNA methylation at this locus. A more detailed overview of the qualitative version is shown in **FIG. 4**.

[**0069**] Detection of methylation in the embodiment of Application D, as in other embodiments herein, is based on amplification-mediated displacement of the probe. In theory, the process of probe displacement might be designed to leave the probe intact, or to result in probe digestion. Preferably, as used herein, displacement of the probe occurs by digestion of the probe during amplification. During the extension phase of the PCR cycle, the fluorescent hybridization probe is cleaved by the 5' to 3' nucleolytic activity of the DNA polymerase. On cleavage of the probe, the reporter moiety emission is no longer transferred efficiently to the quenching moiety, resulting in an increase of the reporter moiety fluorescent-emission spectrum at 518 nm. The fluorescent intensity of the quenching moiety (e.g., TAMRA), changes very little over the course of the PCR amplification. Several factors my influence the efficiency of TaqMan® PCR reactions including: magnesium and salt concentrations; reaction conditions (time and temperature); primer sequences; and PCR target size (i.e., amplicon size) and composition. Optimization of these factors to produce the optimum fluorescence intensity for a given genomic locus is obvious to one skilled in the art of PCR, and preferred conditions are further illustrated in the "Examples" herein. The amnplicon may range in size from 50 to 8,000 base pairs, or larger, but may be smaller. Typically, the amplicon is from 100 to 1000 base pairs, and preferably is from 100 to 500 base pairs. Preferably, the reactions are monitored in real time by performing PCR amplification using 96-well optical trays and caps, and using a sequence detector (ABI Prism) to allow measurement of the fluorescent spectra of all 96 wells of the thermal cycler continuously during the PCR amplification. Preferably, process D is run in combination with the process A **(FIG. 3)** to provide controls for the amount of input nucleic acid, and to normalize data from tray to tray.

[**0070**] Application C.

[**0071**] The inventive process can be modified to avoid sequence discrimination at the PCR product detection level. Thus, in an additional qualitative process embodiment **(FIG.**

3, Application C), just the primers are designed to cover CpG dinucleotides, and sequence discrimination occurs solely at the level of amplification. Preferably, the probe used in this embodiment is still a TaqMan® probe, but is designed so as not to overlap any CpG sequences present in the original, unmodified nucleic acid. The embodiment of Application C represents a high-throughput, fluorescence-based real-time version of MSP technology, wherein a substantial improvement has been attained by reducing the time required for detection of methylated CpG sequences. Preferably, the reactions are monitored in real time by performing PCR amplification using 96-well optical trays and caps, and using a sequence detector (ABI Prism) to allow measurement of the fluorescent spectra of all 96 wells of the thermal cylcer continuously during the PCR amplification. Preferably, process C is run in combination with process A to provide controls for the amount of input nucleic acid, and to normalize data from tray to tray.

[0072] Application B.

[0073] The inventive process can be also be modified to avoid sequence discrimination at the PCR amplification level (**FIG. 3, A** and B). In a quantitative process embodiment (**FIG. 3**, Application B), just the probe is designed to cover CpG dinucleotides, and sequence discrimination occurs solely at the level of probe hybridization. Preferably, TaqMan® probes are used. In this version, sequence variants resulting from the bisulfite conversion step are amplified with equal efficiency; as long as there is no inherent amplification bias (Warnecke et al., *Nucleic Acids Res.* 25:4422-4426, 1997). Design of separate probes for each of the different sequence variants associated with a particular methylation pattern (e.g., 2^3=8 probes in the case of 3 CpGs) would allow a quantitative determination of the relative prevalence of each sequence permutation in the mixed pool of PCR products. Preferably, the reactions are monitored in real time by performing PCR amplification using 96-well optical trays and caps, and using a sequence detector (ABI Prism) to allow measurement of the fluorescent spectra of all 96 wells of the thermal cylcer continuously during the PCR amplification. Preferably, process B is run in combination with process A to provide controls for the amount of input nucleic acid, and to normalize data from tray to tray.

[0074] Application A.

[0075] Process A (**FIG. 3**) does not, in itself, provide for methylated-CpG sequence discrimination at either the amplification or detection levels, but supports and validates the other three applications by providing control reactions for the amount of input DNA, and to normalize data from tray to tray. Thus, if neither the primers, nor the probe overlie any CpG dinucleotides, then the reaction represents unbiased amplification and measurement of amplification using fluorescent-based quantitative real-time PCR serves as a control for the amount of input DNA (**FIG. 3**, Application A). Preferably, process A not only lacks CpG dinucleotides in the primers and probe(s), but also does not contain any CpGs within the amplicon at all to avoid any differential effects of the bisulfite treatment on the amplification process. Preferably, the amplicon for process A is a region of DNA that is not frequently subject to copy number alterations, such as gene amplification or deletion.

[0076] Results obtained with the qualitative version of the technology are described in the examples below. Dozens of human tumor samples have been analyzed using this technology with excellent results. High-throughput using a TaqMan® machine allowed performance of 1100 analyses in three days with one TaqMan® machine.

EXAMPLE 1

[0077] An initial experiment was performed to validate the inventive strategy for assessment of the methylation status of CpG islands in genomic DNA. This example shows a comparison between human sperm DNA (known to be highly unmethylated) and HCT116 DNA (from a human colorectal cell line, known to be highly methylated at many CpG sites) with respect to the methylation status of specific, hypermethylatable CpG islands in four different genes. COBRA (combined bisulfite restriction analysis; Xiong and Laird, *Nucleic Acids Res.* 25:2532-2534, 1997) was used as an independent measure of methylation status.

[0078] DNA Isolation and Bisulfite Treatment.

[0079] Briefly, genomic DNA was isolated from human sperm or HCT116 cells by the standard method of proteinase K digestion and phenol-chloroform extraction (Wolf et al., *Am. J Hum Genet.* 51:478-485, 1992). The DNA was then treated with sodium bisulfite by initially denaturing in 0.2 M NaOH, followed by addition of sodium bisulfite and hydroquinone (to final concentrations of 3.1M, and 0.5M, respectively), incubation for 16 h. at 55° C., desalting (DNA Clean-Up System; Promega), desulfonation by 0.3M NaOH, and final ethanol precipitation. (Xiong and Laird, supra, citing Sadri and Hornsby, *Nucleic Acids Res.* 24:5058-5059, 1996; see also Frommer et al., *Proc. Natl. Acad. Sci. USA* 89:1827-1831, 1992). After bisulfite treatment, the DNA was subjected either to COBRA analysis as previously described (Xiong and Laird, supra), or to the inventive amplification process using fluorescence-based real-time quantitative PCR (Heid et al., *Genome Res.* 6:986-994,-1996; Gibson et al., *Genome-Res.* 6:995-1001, 1996).

[0080] COBRA and MsSNuPE Reactions.

[0081] ESR1 and APC genes were analyzed using COBRA (Combined Bisulfite Restriction Analysis). For COBRA analysis, methylation-dependent sequence differences were introduced into the genomic DNA by standard bisulfite treatment according to the procedure described by Frommer et al (*Proc. Natl. Acad. Sci. USA* 89:1827-1831, 1992) (1 ug of salmon sperm DNA was added as a carrier before the genomic DNA was treated with sodium bisulfite). PCR amplification of the bisulfite converted DNA was performed using primers specific for the interested CpG islands, followed by restriction endonuclease digestion, gel electrophoresis, and detection using specific, labeled hybridization probes. The forward and reverse primer sets used for the ESR1 and APC genes are: TCCTAAAACTACACTTACTCC [SEQ ID NO. 35], GGTTATTTGGAAAAAGAGTATAG [SEQ ID NO. 36] (ESR1 promoter); and AGAGAGAAGTAGTTGTGTTAAT [SEQ ID NO. 37], ACTACACCAATACAACCACAT [SEQ ID NO. 38] (APC promoter), respectively. PCR products of ESR1 were digested by restriction endonuleases TaqI and BstUI, while the products from APC were digested by Taq I and SfaN I, to measure methylation of 3 CpG sies for APC and 4 CpG sites for ESR1. The digested PCR products were electrophoresed on denaturing polyacrylamide gel and transferred to nylon membrane (Zetabind; American Bioanalytical) by

electroblotting. The membranes were hybridized by a 5'-end labeled oligonucleotide to visualize both digested and undigested DNA fragments of interest. The probes used are as follows: ESR1, AAACCAAAACTC [SEQ ID NO. 39]; and APC, CCCACACCCAACCAAT [SEQ ID NO. 40]. Quantitation was performed with the Phosphoimager 445SI (Molecular Dynamics). Calculations were performed in Microsoft Excel. The level of DNA methylation at the investigated CpG sites was determined by calculating the percentage of the digested PCR fragments (Xiong and Laird, supra).

[0082] MLH1 and CDKN2A were analyzed using MsSNuPE (Methylation-sensitive Single Nucleotide Primer Extension Assay), performed as described by Gonzalgo and Jones (*Nucleic Acids Res.* 25:2529-2531). PCR amplification of the bisulfite converted DNA was performed using primers specific for the interested CpG islands, and detection was performed using additional specific primers (extension probes). The forward and reverse primer sets used for the MLH1 and CDKN2A genes are: GGAGGTTATAAGAGTAGGGTTAA [SEQ ID NO. 41], CCAACCAATAAAAACAAAATACC [SEQ ID NO. 42] (MLH1 promoter); GTAGGTGGGGAGGAGTTTAGTT [SEQ ID NO. 43], TCTAATAACCAACCAACCCCTCC [SEQ ID NO. 44] (CDKN2A promoter); and TTGTATTATTTTGTTTTTTTTGGTAGG [SEQ ID NO. 45], CAACTTCTCAAATCATCAATCCTCAC [SEQ ID NO. 46] (CDKN2A Exon 2), respectively. The MsSNuPE extension probes are located immediately 5' of the CpG to be analyzed, and the sequences are: TTTAGTAGAGGTATATAAGTT [SEQ ID NO. 47], TAAGGGGAGAGGAGGAGTTTGAGAAG [SEQ ID NO. 48] (MLH1 promoter sites 1 and 2, respectively); TTTGAGGGATAGGGT [SEQ ID NO. 49], TTTTAGGGGTGTTATATT [SEQ ID NO. 50], TTTTTTTGTTTGGAAAGATAT [SEQ ID NO. 51] (promoter sites 1, 2, and 3, respectively); and GTTGGTGGTGTTGTAT [SEQ ID NO. 52], AGGTTATGATGATGGGTAG [SEQ ID NO. 53], TATTAGAGGTAGTAATTATGTT [SEQ ID NO. 54] Exon2 sites 1, 2, and 3, respectively). A pair of reactions was set up for each sample using either 32p-dCTP or 32p-dTTP for single nucleotide extension. The extended MsSNuPE primers (probes) were separated by denaturing polyacrylamide gel. Quantitation was performed using the Phosphoimager.

[0083] Inventive Methylation Analysis.

[0084] Bisulfite-converted genomic DNA was amplified using locus-specific PCR primers flanking an oligonucleotide probe with a 5' fluorescent reporter dye (6FAM) and a 3' quencher dye (TAMRA) (Livak et al., PCR *Methods Appl.* 4:357-362, 1995) (primers and probes used for the methylation analyses are listed under "Genes, MethyLight Primers and Probe Sequences" herein, infra). In this example, the forward and reverse primers and the corresponding fluorogenic probes were designed to discriminate between either fully methylated or fully unmethylated molecules of bisulfite-converted DNA (see discussion of primer design under "Detailed Description of the Invention, Process D" herein). Primers and a probe were also designed for a stretch of the MYOD1 gene (Myogenic Differentiation Gene), completely devoid of CpG dinucleotides as a control reaction for the amount of input DNA. Parallel reactions were performed using the inventive process with the methylated and unmethylated (D), or control oligos (A) on the bisulfite-treated sperm and HCT116 DNA samples. The values obtained for the methylated and unmethylated reactions were normalized to the values for the MYOD1 control reactions to give the ratios shown in Table 1 (below).

[0085] In a TaqMan® protocol, the 5' to 3' nuclease activity of Taq DNA polymerase cleaved the probe and released the reporter, whose fluorescence was detected by the laser detector of the ABI Prism 7700 Sequence Detection System (Perkin-Elmer, Foster City, Calif.). After crossing a fluorescence detection threshold, the PCR amplification resulted in a fluorescent signal proportional to the amount of PCR product generated. Initial template quantity can be derived from the cycle number at which the fluorescent signal crosses a threshold in the exponential phase of the PCR reaction. Several reference samples were included on each assay plate to verify plate-to-plate consistency. Plates were normalized to each other using these reference samples. The PCR amplification was performed using a 96-well optical tray and caps with a final reaction mixture of 25 μl consisting of 600 nM each primer, 200 nM probe, 200 μM each dATP, dCTP, dGTP, 400 μM dUTP, 5.5 mM $MgCl_2$, 1× TaqMan® Buffer A containing a reference dye, and bisulfite-converted DNA or unconverted DNA at the following conditions: 50° C. for 2 min, 95° C. for 10 min, followed by 40 cycles at 95° C. for 15 s and 60° C. for 1 min.

[0086] Genes, MethyLight Primers and Probe Sequences.

[0087] Four human genes were chosen for analysis: (1) APC (adenomatous polyposis coli) (Hiltunen et al., *Int. J. Cancer* 70:644-648, 1997); (2) ESR1 (estrogen receptor) (Issa et al., *Nature Genet.* 7:536-40, 1994); (3) CDKN2A (p16) (Ahuja, *Cancer Res.* 57:3370-3374, 1997); and (4) hMLH1 (mismatch repair) (Herman et al., *Proc. Natl. Acad. Sci. USA.* 95:6870-6875, 1998; Veigl et al., *Proc. Natl. Acad. Sci. USA.* 95:8698-8702, 1998). These genes were chosen because they contain hypermethylatable CpG islands that are known to undergo de novo methylation in human colorectal tissue in all normal and tumor samples. The human APC gene, for example, has been linked to the development of colorectal cancer, and CpG sites in the regulatory sequences of the gene are known to be distinctly more methylated in colon carcinomas, but not in premalignant adenomas; relative to normal colonic mucosa (Hiltunen et al., supra). The human ESR gene contains a CpG island at its 5' end, which becomes increasingly methylated in colorectal mucosa with age and is heavily methylated in all human colorectal tumors analyzed (Issa et al., supra). Hypermethylation of promoter-associated CpG islands of the CDKN2A (p16) gene has been found in 60% of colorectal cancers showing microsatellite instability (MI) due to defects in one of several base mismatch repair genes (Ahuja et al., supra). The mismatch repair gene MLH1 plays a pivotal role in the development of sporadic cases of mismatch repair-deficient colorectal tumors (Thibodeau et al., Science 260:816-819, 1993). It has been reported that MLH1 can become transcriptionally silenced by DNA hypermethylation of its promoter region, leading to microsatellite instability (MSI) (Kane et al., *Cancer Res.* 57:808-811, 1997; Ahuja et al., supra; Cunningham et al., Cancer Res. 58:3455-3460, 1998; Herman et al., supra; Veigl et al., supra).

[0088] Five sets of PCR primers and probes, designed specifically for bisulfite converted DNA sequences, were

used: (1) a set representing fully methylated and fully unmethylated DNA for the ESR1 gene; (2) a fully methylated set for the MLH1 gene; (3) a fully methylated and fully unmethylated set for the APC gene; and (4) a fully methylated and fully unmethylated set for the CDKN2A (p16) gene; and (5) an internal reference set for the MYOD1 gene to control for input DNA. The methylated and unmethylated primers and corresponding probes were designed to overlap 1 to 5 potential CpG dinucleotides sites. The MYOD1 internal reference primers and probe were designed to cover a region of the MYOD1 gene completely devoid of any CpG dinucleotides to allow for unbiased PCR amplification of the genomic DNA, regardless of methylation status. As indicated above, parallel TaqMan® PCR reactions were performed with primers specific for the bisulfite-converted methylated and/or unmethylated gene sequences and with the MYOD1 reference primers. The primer and probe sequences are listed below. In all cases, the first primer listed is the forward PCR primer, the second is the TaqMan® probe, and the third is the reverse PCR primer. ESR1 methylated (GGCGTTCGTTTTGGGATTG [SEQ ID NO. 1], 6FAM 5'-CGATAAAACCGAACGACCCGACGA-3' TAMRA [SEQ ID NO. 2], GCCGACACGCGAACTCTAA [SEQ ID NO. 3]); ESR1 unmethylated (ACACATATCCCACCAACACACAA [SEQ ID NO. 4], 6FAM 5'-CAACCCTACCCCAAAAACCTACAAATCCAA-3'TAMRA [SEQ ID NO. 5], AGGAGTTGGTGGAGGGTGTTT [SEQ ID NO. 6]); MLH1 methylated (CTATCGCCGCCTCATCGT [SEQ ID NO. 7], 6FAM 5'-CGCGACGTCAAACGCCACTACG-3' TAMRA [SEQ ID NO. 8], CGTTATATATCGTTCGTAGTATTCGTGTTT [SEQ ID NO. 9]); APC methylated (TTATATGTCGGTTACGTGCGTTTATAT [SEQ ID NO. 10], 6FAM 5'-CCCGTCGAAAACCCGCCGATTA-3' TAMRA [SEQ ID NO. 11], GAACCAAAACGCTCCCCAT [SEQ ID NO. 12]); APC unmethylated (GGGTTGTGAGGGTATATTTTTGAGG [SEQ ID NO. 13], 6FAM 5'-CCCACCCAACCACACAACCTACCTAACC-3' TAMRA [SEQ ID NO. 14], CCAACCCACACTCCACAATAAA [SEQ ID NO. 15]); CDKN2A methylated (AACAACGTCCGCACCTCCT [SEQ ID NO. 16], 6FAM 5'-ACCCGACCCCGAACCGCG-3'TAMRA [SEQ ID NO. 17], TGGAATTTTCGGTTGATTGGTT [SEQ ID NO. 18]); CDKN2A unmethylated (CAACCAATCAACCAAAAATTCCAT [SEQ ID NO. 19], 6FAM 5'-CCACCACCCACTATCTACTCTCCCCCTC-3'TAMRA [SEQ ID NO. 20], GGTGGATTGTGTGTGTTTGGTG [SEQ ID NO. 21]); and MYOD1, (CCAACTCCAAATCCCCTCTCTAT [SEQ ID NO. 22], 6FAM 5'-TCCCTTCCTATTCCTAAATCCAACCTAAATACCTCC-3' TAMRA [SEQ ID NO. 23], TGATTAATTTAGATTGGGTTTAGAGAAGGA [SEQ ID NO. 24]).

[**0089**] Tables 1 and 2 shows the results of the analysis of human sperm and HCT116 DNAs for methylation status of the CpG islands within the four genes; APC, ESR1, CDKN2A (p16), and hMLH1. The results are expressed as ratios between the methylated and unmethylated reactions and a control reaction (MYOD1). Table 1 shows that sperm DNA yielded a positive ratio only with the "unmethylated" primers and probe; consistent with the known unmethylated status of sperm DNA, and consistent with the percent methylation values determined by COBRA analysis. That is, priming on the bisulfite-treated DNA occurred from regions that contained unmethylated cytosine in CpG sequences in the corresponding genomic DNA, and hence were deaminated (converted to uracil) by bisulfite treatment.

TABLE 1

Technique	COBRA or Ms-SNuPE	Methylated Reaction*	Unmethylated Reaction*
GENE			
APC	0%	0	49
ESR1	0%	0	62
CDKN2A	0%**	0	52
hMLH1	ND	0	ND

*The values do not represent percentages, but values in an arbitrary unit that can be compared quantitatively between different DNA samples for the same reaction, after normalization with a control gene.
**Based on Ms-SNuPE.

[**0090**] Table 2 shows the results of an analysis of HCT116 DNA for methylation status of the CpG islands within the four genes; APC, ESR1, CDKN2A (p16), and hMLH1. The results are expressed as ratios between the methylation-specific reactions and a control reaction (MYOD1). For the ESR gene, a positive ratio was obtained only with the "methylated" primers and probe; consistent with the known methylated status of HCT116 DNA, and the COBRA analysis. For the CDKN2A gene, HCT116 DNA yielded positive ratios with both the "methylated" and "unmethylated" primers and probe; consistent with the known methylated status of HCT116 DNA, and with the COBRA analysis that indicates only partial methylation of this region of the gene. By contrast, the APC gene gave positive results only with the unmethylated reaction. However, this is entirely consistent with the COBRA analysis, and indicates that this APC gene region is unmethylated in HCT116 DNA. This may indicate that the methylation state of this particular APC gene regulatory region in the DNA from the HCT116 cell line is more like that of normal colonic mucosa or premalignant adenomas rather than that of colon carcinomas (known to be distinctly more methylated).

TABLE 2

Technique	COBRA and/or Ms-SNuPE	Methylated Reaction*	Unmethylated Reaction*
GENE			
APC	2%	0	81
ESR1	99%	36	0
CDKN2A	38%**	222	26
hMLH1	ND	0	ND

*The values do not represent percentages, but values in an arbitrary unit that can be compared quantitatively between different DNA samples for the same reaction, after normalization with a control gene.
**Based on Ms-SNuPE.

EXAMPLE 2

[**0091**] This example is a comparison of the inventive process (A and D in **FIG. 3**) with an independent COBRA method (See "Methods," above) to determine the methylation status of a CpG island associated with the estrogen receptor (ESR1) gene in the human colorectal cell line HCT116 and in human sperm DNA. This CpG island has been reported to be highly methylated in HCT116 and unmethylated in human sperm DNA (Xiong and Laird, supra; Issa et al., supra). The COBRA analysis, is described above. Two TaqI sites within this CpG island confirmed this,

showing a lack of methylation in the sperm DNA and nearly complete methylation in HCT116 DNA (**FIG. 5A**). Additionally, results using bisulfite-treated and untreated DNA were compared.

[**0092**] For an analysis, fully "methylated" and fully "unmethylated" ESR1, and control MYOD1 primers and probes were designed as described above under "Example 1." Three separate reactions using either the "methylated,""unmethylated" or control oligos on both sperm and HCT116 DNA were performed. As in Example 1, above, the values obtained for the methylated and unmethylated reactions were normalized to the values for the MYOD1 control reactions to give the ratios shown in **FIG. 5B**. Sperm DNA yielded a positive ratio only with the unmethylated primers and probe, consistent with its unmethylated status. In contrast, HCT116 DNA, with predominantly methylated ESR1 alleles, generated a positive ratio only in the methylated reaction (**FIG. 5B**). Both the sperm and HCT116 DNA yielded positive values in the MYOD1 reactions, indicating that there was sufficient input DNA for each sample. As expected, the non-bisulfite converted DNA with either the methylated or unmethylated oligonucleotides (**FIG. 5B**) was not amplified. These results are consistent with the COBRA findings (**FIG. 5A**), suggesting that the inventive assay can discriminate between the methylated and unmethylated alleles of the ESR1 gene. In addition, the reactions are specific to bisulfite-converted DNA, which precludes the generation of false positive results due to incomplete bisulfite conversion.

EXAMPLE 3

[**0093**] This example determined specificity of the inventive primers and probes. **FIG. 6** shows a test of all possible combinations of primers and probes to further examine the specificity of the methylated and unmethylated oligonucleotides on DNAs of known methylation status. Eight different combinations of the ESR1 "methylated" and "unmethylated" forward and reverse primers and probe (as described above in "Example 1") were tested in different combinations in inventive assays on sperm and HCT116 DNA in duplicate. The assays were performed as described above in Example 1. Panel A (**FIG. 6**) shows the nomenclature used for the combinations of the ESR1 oligos. "U" refers to the oligo sequence that anneals with bisulfite-converted unmethylated DNA, while "M" refers to the methylated version. Position 1 indicates the forward PCR primer, position 2 the probe, and position 3 the reverse primer. The combinations used for the eight reactions are shown below each pair of bars, representing duplicate experiments. The results are expressed as ratios between the ESR1 values and the MYOD1 control values. Panel B represents an analysis of human sperm DNA. Panel C represents an analysis of DNA obtained from the human colorectal cancer cell line HCT116.

[**0094**] Only the fully unmethylated (reaction 1) or fully methylated combinations (reaction 8) resulted in a positive reaction for the sperm and HCT116, respectively. The other combinations were negative, indicating that the PCR conditions do not allow for weak annealing of the mismatched oligonucleotides. This selectivity indicates that the inventive process can discriminate between fully methylated or unmethylated alleles with a high degree of specificity.

EXAMPLE 4

[**0095**] This example shows that the inventive process is reproducible. **FIG. 7** illustrates an analysis of the methylation status of the ESR1 locus in DNA samples derived from a primary colorectal adenocarcinoma and matched normal mucosa derived from the same patient (samples 10N and 10T in **FIG. 8**) in order to study a heterogeneous population of methylated and unmethylated alleles. The colorectal tissue samples were collected as described in Example 5, below. In addition, the reproducibility of the inventive process was tested by performing eight independent reactions for each assay. The results for the ESR1 reactions and for the MYOD1 control reaction represent raw absolute values obtained for these reactions, rather than ratios, so that the standard errors of the individual reactions can be evaluated. The values have been plate-normalized, but not corrected for input DNA. The bars indicate the mean values obtained for the eight separate reactions. The error bars represent the standard error of the mean.

[**0096**] **FIG. 7** shows that the mean value for the methylated reaction was higher in the tumor compared to the normal tissue whereas the unmethylated reaction showed the opposite result. The standard errors observed for the eight independent measurements were relatively modest and were comparable to those reported for other studies utilizing TaqMan® technology (Fink et al., *Nature Med.* 4:1329-1333, 1998). Some of the variability of the inventive process may have been a result of stochastic PCR amplification (PCR bias), which can occur at low template concentrations. (Warnecke et al., *Nucleic Acids Res.* 25:4422-4426,1997). In summary, these results indicate that the inventive process can yield reproducible results for complex, heterogeneous DNA samples.

EXAMPLE 5

[**0097**] This example shows a comparison of MLHI Expression, microsatellite instability and MLH1 promoter methylation in 25 matched-paired human colorectal samples. The main benefit of the inventive process is the ability to rapidly screen human tumors for the methylation state of a particular locus. In addition, the analysis of DNA methylation as a surrogate marker for gene expression is a novel way to obtain clinically useful information about tumors. We tested the utility of the inventive process by interrogating the methylation status of the MLH1 promoter. The mismatch repair gene MLH1 plays a pivotal role in the development of sporadic cases of mismatch repair-deficient colorectal tumors (Thibodeau et al., *Science* 260:816-819, 1993). It has been reported that MLH1 can become transcriptionally silenced by DNA hypermethylation of its promoter region, leading to microsatellite instability (MSI) (Kane et al., *Cancer Res* 57:808-811, 1997;,Ahuja et al., *Cancer Res* 57:3370-3374, 1997; Cunningham et al., *Cancer Res.* 58:3455-3460, 1998; Herman, J. G. et al., *Proc. Natl. Acad. Sci. USA* 95:6870-6875, 1998; Veigl et al., *Proc. Natl. Acad. Sci. USA* 95:8698-8702, 1998).

[**0098**] Using the high-throughput inventive process, as described in Example 1 Application D, 50 samples consisting of 25 matched pairs of human colorectal adenocarcinomas and normal mucosa were analyzed for the methylation status of the MLH1 CpG island. Quantitative RT-PCR (TaqMan®) analyses of the expression levels of MLH1

normalized to ACTB (β-actin) was investigated. Furthermore, the microsatellite instability (MSI) status of each sample was analyzed by PCR of the BAT25 and BAT26 loci (Parsons et al., *Cancer Res.* 55:5548-5550, 1995). The twenty-five paired tumor and normal mucosal tissue samples were obtained from 25 patients with primary colorectal adenocarcinoma. The patients comprised 16 males and 9 females, ranging in age from 39-88 years, with a mean age of 68.8. The mucosal distance from tumor to normal specimens was between 10 and 20 cm. Approximately 2 grams of the surgically removed tissue was immediately frozen in liquid nitrogen and stored at −80° C. until RNA and DNA isolation.

[**0099**] Quantitative RT-PCR and Microsatellite Instability Analysis.

[**0100**] The quantitation of mRNA levels was carried out using real-time fluorescence detection. The TaqMan® reactions were performed as described above for the assay, but with the addition of 1U AmpErase uracil N-glycosylase). After RNA isolation, cDNA was prepared from each sample as previously described (Bender et al., *Cancer Res* 58:95-101, 1998). Briefly, RNA was isolated by lysing tissue in buffer containing quanidine isothiocyanate (4M), N-lauryl sarcosine (0.5%), sodium citrate (25 mM), and 2-mercaptoethanol (0.1M), followed by standard phenol-chloroform extraction, and precipitation in 50% isopropanol/50% lysis buffer. To prepare cDNA, RNA samples were reverse-transcribed using random hexamers, deoxynucleotide triphosphates, and Superscript II® reverse transcriptase (Life Technologies, Inc., Palo Alto, Calif.). The resulting cDNA was then amplified with primers specific for MLH1 and ACTB. Contamination of the RNA samples by genomic DNA was excluded by analysis of all RNA samples without prior cDNA conversion. Relative gene expression was determined based on the threshold cycles (number of PCR cycles required for detection with-a specific probe) of the MLH1 gene and of the internal reference gene ACTB. The forward primer, probe and reverse primer sequences of the ACTB and MLH1 genes are: ACTB (TGAGCGCGGCTACAGCTT [SEQ ID NO. 25], 6FAM5'-ACCACCACGGCCGAGCGG-3'TAMRA [SEQ ID NO. 26], CCTTAATGTCACACACGATT [SEQ ID NO. 27]); and MLH1 (GTTCTCCGGGAGATGTTGCATA [SEQ ID NO. 28], 6FAM5'-CCTCAGTGGGCCTTGGCACAGC-3'TAMRA [SEQ ID NO. 29], TGGTGGTGTTGAGAAGGTATAACTTG [SEQ ID NO. 30]).

[**0101**] Alterations of numerous polyadenine ("pA") sequences, distributed widely throughout the genome, is a useful characteristic to define tumors with microsatellite instability (Ionov et al., *Nature* 363-558-561, 1993). Microsatellite instability (MSI) was determined by PCR and sequence analysis of the BAT25 (25-base pair pA tract from an intron of the c-kit oncogene) and BAT26 (26-base pair pA tract from an intron of the mismatch repair gene hMSH2) loci as previously described (Parsons et al., *Cancer Res* 55:5548-5550, 1995). Briefly, segments the BAT25 and BAT26 loci were amplified for 30 cycles using one ^{32}P-labeled primer and one unlabeled primer for each locus. Reactions were resolved on urea-formamide gels and exposed to film. The forward and reverse primers that were used for the amplification of BAT25 and BAT26 were: BAT25 (TCGCCTCCAAGAATGTAAGT [SEQ ID NO. 31], TCTGCATTTTAACTATGGCTC [SEQ ID NO. 32]); and BAT26 (TGACTACTTTTGACTTCAGCC [SEQ ID NO. 33], AACCATTCAACATTTTTAACCC [SEQ ID NO. 34]).

[**0102**] **FIG. 8** shows the correlation between MLH1 gene expression, MSI status and promoter methylation of MLH1, as determined by the inventive process. The upper chart shows the MLH1 expression levels measured by quantitative, real time RT-PCR (TaqMan®) in matched normal (hatched bars) and tumor (solid black bars) colorectal samples. The expression levels are displayed as a ratio between MLH1 and ACTB measurements. Microsatellite instability status (MSI) is indicated by the circles located between the two charts. A black circle denotes MSI positivity, while an open circle indicates that the sample is MSI negative, as determined by analysis of the BAT25 and BAT26 loci. The lower chart shows the methylation status of the MLH1 locus as determined by inventive process. The methylation levels are represented as the ratio between the MLH1 methylated reaction and the MYOD1 reaction.

[**0103**] Four colorectal tumors had significantly elevated methylation levels compared to the corresponding normal tissue. One of these (tumor 17) exhibited a particularly high degree of MLH1 methylation, as scored by the inventive process. Tumor 17 was the only sample that was both MSI positive (black circle) and showed transcriptional silencing of MLH1. The remaining methylated tumors expressed MLH1 at modest levels and were MSI negative (white circle). These results show that MLH1 was biallelically methylated in tumor 17, resulting in epigenetic silencing and consequent microsatellite instability, whereas the other tumors showed lesser degrees of MLH1 promoter hypermethylation and could have just one methylated allele, allowing expression from the unaltered allele. Accordingly, the inventive process was capable of rapidly generating significant biological information, such as promoter CpG island hypermethylation in human tumors, which is associated with the transcriptional silencing of genes relevant to the cancer process.

[**0104**] COBRA and MsSNuPE Reactions.

[**0105**] ESR1 and ABC genes were analyzed using COBRA (Combined Bisulfite Restriction Analysis). For COBRA analysis, methylation-dependent sequence differences were introduced into the genomic DNA by standard bisulfite treatment according to the procedure described by Frommer et al (*Proc. Natl. Acad. Sci. USA* 89:1827-1831, 1992) (1 ug of salmon sperm DNA was added as a carrier before the genomic DNA was treated with sodium bisulfite). PCR amplification of the bisulfite converted DNA was performed using primers specific for the interested CpG islands, followed by restriction endonuclease digestion, gel electrophoresis, and detection using specific, labeled hybridization probes. The forward and reverse primer sets used for the ESR1 and APC genes are: TCCTAAAACTACACTTACTCC [SEQ ID NO. 35], GGTTATTTGGAAAAAGAGTATAG [SEQ ID NO. 36] (ESR1 promoter); and AGAGAGAAGTAGTTGTGTTAAT [SEQ ID NO. 37], ACTACACCAATACAACCACAT [SEQ ID NO. 38] (APC promoter), respectively. PCR products of ESR1 were digested by restriction endonuleases TaqI and BstUI, while the products from APC were digested by Taq I and SfaN I, to measure methylation of 3 CpG sies for APC and 4 CpG sites for ESR1. The digested PCR products were electro-

phoresed on denaturing polyacrylamide gel and transferred to nylon membrane (Zetabind; American Bioanalytical) by electroblotting. The membranes were hybridized by a 5'-end labeled oligonucleotide to visualize both digested and undigested DNA fragments of interest. The probes used are as follows: ESR1, AAACCAAAACTC [SEQ ID NO. 39]; and APC, CCCACACCCAACCAAT [SEQ ID NO. 40]. Quantitation was performed with the Phosphoimager 445SI (Molecular Dynamics). Calculations were performed in Microsoft Excel. The level of DNA methylation at the investigated CpG sites was determined by calculating the percentage of the digested PCR fragments (Xiong and Laird, supra).

[**0106**] A MLH1 and CDKN2A were analyzed using MsSNuPE (Methylation-sensitive Single Nucleotide Primer Extension Assay), performed as described by Gonzalgo and Jones (*Nucleic Acids Res.* 25:2529-2531). PCR amplification of the bisulfite converted DNA was performed using primers specific for the interested CpG islands, and detection was performed using additional specific primers (extension probes). The forward and reverse primer sets used for the MLH1 and CDKN2A genes are: GGAGGTTATAAGAGTAGGGTTAA [SEQ ID NO. 41], CCAACCAATAAAAACAAAAATACC [SEQ ID NO. 42] (MLH1 promoter); GTAGGTGGGGAGGAGTTTAGTT [SEQ ID NO. 43], TCTAATAACCAACCAACCCCTCC [SEQ ID NO. 44] (CDKN2A promoter); and TTGTATTATTTTGTTTTTTTGGTAGG [SEQ ID NO. 45], CAACTTCTCAAATCATCAATCCTCAC [SEQ ID NO. 46] (CDKN2A Exon 2), respectively. The MsSNuPE extension probes are located immediately 5' of the CpG to be analyzed, and the sequences are: TTTAGTAGAGGTATATAAGTT [SEQ ID NO. 47], TAAGGGGAGAGGAGGAGTTTGAGAAG [SEQ ID NO. 48] (MLH1 promoter sites 1 and 2, respectively), TTTGAGGGATAGGGT [SEQ ID NO. 49], TTTTAGGGGTGTTATATT [SEQ ID

SEQUENCE LISTING

(1) GENERAL INFORMATION:

(iii) NUMBER OF SEQUENCES: 54

(2) INFORMATION FOR SEQ ID NO: 1:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 19 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA

(iii) HYPOTHETICAL: No

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 1:

GGCGTTCGTT TTGGGATTG 19

(2) INFORMATION FOR SEQ ID NO: 2:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 24 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA

(iii) HYPOTHETICAL: No

(ix) FEATURE:
(A) NAME/KEY: 5' substitution with fluorescent reporter dye 6FAM (2,7-dimethoxy-4,5-dichloro-6-carboxy-fluorescein-phosporamidite-cytosine); 3'substitution with quencher dye TAMRA (6-carboxytetramethylrhodamine).

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 2:

CGATAAAACC GAACGACCCG ACGA 24

(2) INFORMATION FOR SEQ ID NO: 3:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 19 base pairs

-continued

(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA

(iii) HYPOTHETICAL: No

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 3:

GCCGACACGC GAACTCTAA 19

(2) INFORMATION FOR SEQ ID NO: 4:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 23 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA

(iii) HYPOTHETICAL: No

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 4:

ACACATATCC CACCAACACA CAA 23

(2) INFORMATION FOR SEQ ID NO: 5:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 30 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA

(iii) HYPOTHETICAL: No

(ix) FEATURE:
(A) NAME/KEY: 5' substitution with fluorescent reporter dye 6FAM (2,7-dimethoxy-4,5-dichloro-6-carboxy-fluorescein-phosporamidite-cytosine); 3'substitution with quencher dye TAMRA (6-carboxytetramethylrhodamine).

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 5:

CAACCCTACC CCAAAAACCT ACAAATCCAA 30

(2) INFORMATION FOR SEQ ID NO: 6:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 21 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA

(iii) HYPOTHETICAL: No

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 6:

AGGAGTTGGT GGAGGGTGTT T 21

(2) INFORMATION FOR SEQ ID NO: 7:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 18 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single

-continued

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA

(iii) HYPOTHETICAL: No

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 7:

CTATCGCCGC CTCATCGT 18

(2) INFORMATION FOR SEQ ID NO: 8:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 22 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA

(iii) HYPOTHETICAL: No

(ix) FEATURE:
(A) NAME/KEY: 5' substitution with fluorescent reporter dye 6FAM (2,7-dimethoxy-4,5-dichloro-6-carboxy-fluorescein-phosporamidite-cytosine); 3'substitution with quencher dye TAMRA (6-carboxytetramethylrhodamine).

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 8:

CGCGACGTCA AACGCCACTA CG 22

(2) INFORMATION FOR SEQ ID NO: 9:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 30 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA

(iii) HYPOTHETICAL: No

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 9:

CGTTATATAT CGTTCGTAGT ATTCGTGTTT 30

(2) INFORMATION FOR SEQ ID NO: 10:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 27 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA

(iii) HYPOTHETICAL: No

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 10:

TTATATGTCG GTTACGTGCG TTTATAT 27

(2) INFORMATION FOR SEQ ID NO: 11:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 22 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

-continued

(ii) MOLECULE TYPE: DNA

(iii) HYPOTHETICAL: No

(ix) FEATURE:
(A) NAME/KEY: 5' substitution with fluorescent reporter dye 6FAM (2,7-dimethoxy-4,5-dichloro-6-carboxy-fluorescein-phosporamidite-cytosine); 3'substitution with quencher dye TAMRA (6-carboxytetramethylrhodamine).

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 11:

CCCGTCGAAA ACCCGCCGAT TA 22

(2) INFORMATION FOR SEQ ID NO: 12:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 19 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA

(iii) HYPOTHETICAL: No

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 12:

GAACCAAAAC GCTCCCCAT 19

(2) INFORMATION FOR SEQ ID NO: 13:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 25 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA

(iii) HYPOTHETICAL: No

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 13:

GGGTTGTGAG GGTATATTTT TGAGG 25

(2) INFORMATION FOR SEQ ID NO: 14:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 28 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA

(iii) HYPOTHETICAL: No

(ix) FEATURE:
(A) NAME/KEY: 5' substitution with fluorescent reporter dye 6FAM (2,7-dimethoxy-4,5-dichloro-6-carboxy-fluorescein-phosporamidite-cytosine); 3'substitution with quencher dye TAMRA (6-carboxytetramethylrhodamine).

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 14:

CCCACCCAAC CACACAACCT ACCTAACC 28

(2) INFORMATION FOR SEQ ID NO: 15:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 22 base pairs

-continued

(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA

(iii) HYPOTHETICAL: No

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 15:

CCAACCCACA CTCCACAATA AA 22

(2) INFORMATION FOR SEQ ID NO: 16:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 19 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA

(iii) HYPOTHETICAL: No

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 16:

AACAACGTCC GCACCTCCT 19

(2) INFORMATION FOR SEQ ID NO: 17:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 18 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA

(iii) HYPOTHETICAL: No

(ix) FEATURE:
(A) NAME/KEY: 5' substitution with fluorescent reporter dye 6FAM (2,7-dimethoxy-4,5-dichloro-6-carboxy-fluorescein-phosporamidite-cytosine); 3'substitution with quencher dye TAMRA (6-carboxytetramethylrhodamine).

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 17:

ACCCGACCCC GAACCGCG 18

(2) INFORMATION FOR SEQ ID NO: 18:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 22 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA

(iii) HYPOTHETICAL: No

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 18

TGGAATTTTC GGTTGATTGG TT 22

(2) INFORMATION FOR SEQ ID NO: 19:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 24 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single

-continued

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA

(iii) HYPOTHETICAL: No

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 19:

CAACCAATCA ACCAAAAATT CCAT 24

(2) INFORMATION FOR SEQ ID NO: 20

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 28 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA

(iii) HYPOTHETICAL: No

(ix) FEATURE:
(A) NAME/KEY: 5' substitution with fluorescent reporter dye 6FAM(2,7 dimethoxy-4,5-dichloro-6-carboxy-fluorescein-phosporamidite-cytosine); 3'substitution with quencher dye TAMRA (6-carboxytetramethylrhodamine).

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 20:

CCACCACCCA CTATCTACTC TCCCCCTC 28

(2) INFORMATION FOR SEQ ID NO: 21:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 22 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA

(iii) HYPOTHETICAL: No

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 21:

GGTGGATTGT GTGTGTTTGG TG 22

(2) INFORMATION FOR SEQ ID NO: 22:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 23 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA

(iii) HYPOTHETICAL: No

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 22:

CCAACTCCAA ATCCCCTCTC TAT 23

(2) INFORMATION FOR SEQ ID NO: 23:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 36 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

-continued

(ii) MOLECULE TYPE: DNA

(iii) HYPOTHETICAL: No

(ix) FEATURE:
(A) NAME/KEY: 5' substitution with fluorescent reporter dye 6FAM (2,7-dimethoxy-4,5-dichloro-6-carboxy-fluorescein-phosporamidite-cytosine); 3'substitution with quencher dye TAMRA (6-carboxytetramethylrhodamine).

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 23:

TCCCTTCCTA TTCCTAAATC CAACCTAAAT ACCTCC 36

(2) INFORMATION FOR SEQ ID NO: 24:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 30 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA

(iii) HYPOTHETICAL: No

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 24:

TGATTAATTT AGATTGGGTT TAGAGAAGGA 30

(2) INFORMATION FOR SEQ ID NO: 25:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 18 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA

(iii) HYPOTHETICAL: No

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 25:

TGAGCGCGGC TACAGCTT 18

(2) INFORMATION FOR SEQ ID NO: 26:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 18 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA

(iii) HYPOTHETICAL: No

(ix) FEATURE:
(A) NAME/KEY: 5' substitution with fluorescent reporter dye 6FAM (2,7-dimethoxy-4,5-dichloro-6-carboxy-fluorescein-phosporamidite-cytosine); 3'substitution with quencher dye TAMRA (6-carboxytetramethylrhodamine).

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 26:

ACCACCACGG CCGAGCGG 18

(2) INFORMATION FOR SEQ ID NO: 27:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 20 base pairs

-continued

(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA

(iii) HYPOTHETICAL: No

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 27:

CCTTAATGTC ACACACGATT 20

(2) INFORMATION FOR SEQ ID NO: 28:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 22 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA

(iii) HYPOTHETICAL: No

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 28:

GTTCTCCGGG AGATGTTGCA TA 22

(2) INFORMATION FOR SEQ ID NO: 29:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 22 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA

(iii) HYPOTHETICAL: No

(ix) FEATURE:
(A) NAME/KEY: 5' substitution with fluorescent reporter dye 6FAM (2,7-dimethoxy-4,5-dichloro-6-carboxy-fluorescein-phosporamidite-cytosine); 3'substitution with quencher dye TAMRA (6-carboxytetramethylrhodamine).

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 29:

CCTCAGTGGG CCTTGGCACA GC 22

(2) INFORMATION FOR SEQ ID NO: 30:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 26 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA

(iii) HYPOTHETICAL: No

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 30:

TGGTGGTGTT GAGAAGGTAT AACTTG 26

(2) INFORMATION FOR SEQ ID NO: 31:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 20 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single

-continued

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA

(iii) HYPOTHETICAL: No

(x) PUBLICATION INFORMATION:
(A) AUTHORS: Parsons, et al
(B) TITLE: Microsatellite Instability and Mutations of the Transforming Growth Factor B Type II Receptor Gene in Colorectal Cancer
(C) JOURNAL: Cancer Res.
(D) VOLUME: 55
(F) PAGES: 5548-5550
(G) DATE: 01-DEC-1995

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 31:

TCGCCTCCAA GAATGTAAGT 20

(2) INFORMATION FOR SEQ ID NO: 32:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 21 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA

(iii) HYPOTHETICAL: No

(x) PUBLICATION INFORMATION:
(A) AUTHORS: Parsons, et al
(B) TITLE: Microsatellite Instability and Mutations of the Transforming Growth Factor B Type II Receptor Gene in Colorectal Cancer
(C) JOURNAL: Cancer Res.
(D) VOLUME: 55
(F) PAGES: 5548-5550
(G) DATE: 01-DEC-1995

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 32:

TCTGCATTTT AACTATGGCT C 21

(2) INFORMATION FOR SEQ ID NO: 33:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 21 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA

(iii) HYPOTHETICAL: No

(x) PUBLICATION INFORMATION:
(A) AUTHORS: Parsons, et al
(B) TITLE: Microsatellite Instability and Mutations of the Transforming Growth Factor B Type II Receptor Gene in Colorectal Cancer
(C) JOURNAL: Cancer Res.
(D) VOLUME: 55
(F) PAGES: 5548-5550
(G) DATE: 01-DEC-1995

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 33:

TGACTACTTT TGACTTCAGC C 21

(2) INFORMATION FOR SEQ ID NO: 34:

-continued

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 22 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA

(iii) HYPOTHETICAL: No

(x) PUBLICATION INFORMATION:
(A) AUTHORS: Parsons, et al
(B) TITLE: Microsatellite Instability and Mutations of the Transforming Growth Factor B Type II Receptor Gene in Colorectal Cancer
(C) JOURNAL: Cancer Res.
(D) VOLUME: 55
(F) PAGES: 5548-5550
(G) DATE: 01-DEC-1995

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 34:

AACCATTCAA CATTTTTAAC CC 22

(2) INFORMATION FOR SEQ ID NO: 35:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 21 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA

(iii) HYPOTHETICAL: No

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 35:

TCCTAAAACT ACACTTACTC C 21

(2) INFORMATION FOR SEQ ID NO: 36:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 23 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA

(iii) HYPOTHETICAL: No

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 36:

GGTTATTTGG AAAAAGAGTA TAG 23

(2) INFORMATION FOR SEQ ID NO: 37:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 22 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA

(iii) HYPOTHETICAL: No

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 37:

AGAGAGAAGT AGTTGTGTTA AT 22

-continued

(2) INFORMATION FOR SEQ ID NO: 38:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 21 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA

(iii) HYPOTHETICAL: No

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 38:

ACTACACCAA TACAACCACA T 21

(2) INFORMATION FOR SEQ ID NO: 39:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 12 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA

(iii) HYPOTHETICAL: No

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 39:

AAACCAAAAC TC 12

(2) INFORMATION FOR SEQ ID NO: 40:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 16 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA

(iii) HYPOTHETICAL: No

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 40:

CCCACACCCA ACCAAT 16

(2) INFORMATION FOR SEQ ID NO: 41:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 23 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA

(iii) HYPOTHETICAL: No

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 41:

GGAGGTTATA AGAGTAGGGT TAA 23

(2) INFORMATION FOR SEQ ID NO: 42:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 24 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

-continued

(ii) MOLECULE TYPE: DNA

(iii) HYPOTHETICAL: No

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 42:

CCAACCAATA AAAACAAAAA TACC 24

(2) INFORMATION FOR SEQ ID NO: 43:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 22 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA

(iii) HYPOTHETICAL: No

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 43:

GTAGGTGGGG AGGAGTTTAG TT 22

(2) INFORMATION FOR SEQ ID NO: 44:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 23 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA

(iii) HYPOTHETICAL: No

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 44:

TCTAATAACC AACCAACCCC TCC 23

(2) INFORMATION FOR SEQ ID NO: 45:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 27 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA

(iii) HYPOTHETICAL: No

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 45:

TTGTATTATT TTGTTTTTTT TGGTAGG 27

(2) INFORMATION FOR SEQ ID NO: 46:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 26 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA

(iii) HYPOTHETICAL: No

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 46:

CAACTTCTCA AATCATCAAT CCTCAC 26

-continued

(2) INFORMATION FOR SEQ ID NO: 47:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 21 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA

(iii) HYPOTHETICAL: No

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 47:

TTTAGTAGAG GTATATAAGT T 21

(2) INFORMATION FOR SEQ ID NO: 48:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 26 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA

(iii) HYPOTHETICAL: No

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 48:

TAAGGGGAGA GGAGGAGTTT GAGAAG 26

(2) INFORMATION FOR SEQ ID NO: 49:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 15 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA

(iii) HYPOTHETICAL: No

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 49:

TTTGAGGGAT AGGGT 15

(2) INFORMATION FOR SEQ ID NO: 50:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 18 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA

(iii) HYPOTHETICAL: No

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 50:

TTTTAGGGGT GTTATATT 18

(2) INFORMATION FOR SEQ ID NO: 51:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 21 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single

-continued

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA

(iii) HYPOTHETICAL: No

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 51:

TTTTTTTGTT TGGAAAGATA T 21

(2) INFORMATION FOR SEQ ID NO: 52:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 16 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA

(iii) HYPOTHETICAL: No

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 52:

GTTGGTGGTG TTGTAT 16

(2) INFORMATION FOR SEQ ID NO: 53:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 19 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA

(iii) HYPOTHETICAL: No

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 53:

AGGTTATGAT GATGGGTAG 19

(2) INFORMATION FOR SEQ ID NO: 54:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 22 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA

(iii) HYPOTHETICAL: No

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 54:

TATTAGAGGT AGTAATTATG TT 22

We claim:

1. A method for detecting cytosine methylation and methylated CpG islands within a genomic sample of DNA comprising:

(a) contacting a genomic sample of DNA with a modifying agent that modifies unmethylated cytosine to produce a converted nucleic acid;

(b) amplifying the converted nucleic acid by means of two oligonucleotide primers in the presence or absence of one or a plurality of specific oligonucleotide probes, wherein one or a plurality of oligonucleotide primers and/or the specific probe(s) are capable of distinguishing between unmethylated and methylated nucleic acid; and

(c) detecting the methylated nucleic acid based on amplification-mediated digestion of the probe.

2. The method of claim 1 wherein the amplifying step is a polymerase chain reaction (PCR).

3. The method of claim 1 wherein the modifying agent is bisulfite.

4. The method of claim 1 wherein the converted nucleic acid contains uracil in place of unmethylated cytosine residues present in the unmodified nucleic acid-containing sample.

5. The method of claim 1 wherein the probe further comprises one or a plurality of fluorescence label moieties.

6. The method of claim 5 wherein the amplification and detection step comprises fluorescence-based quantitative PCR.

7. A method for detecting a methylated CpG-containing nucleic acid comprising:

(a) contacting a nucleic acid-containing sample with a modifying agent that modifies unmethylated cytosine to produce a converted nucleic acid;

(b) amplifying the converted nucleic acid in the sample by means of oligonucleotide primers in the presence of a CpG-specific oligonucleotide probe, wherein the CpG-specific probe, but not the primers, distinguish between modified unmethylated and methylated nucleic acid; and

(c) detecting the methylated nucleic acid based upon an amplification-mediated displacement of the CpG-specific probe.

8. The method of claim 7 wherein the amplifying step comprises a polymerase chain reaction (PCR).

9. The method of claim 7 wherein the modifying agent comprises bisulfite.

10. The method of claim 7 wherein the converted nucleic acid contains uracil in place of unmethylated cytosine residues present in the unmodified nucleic acid-containing sample.

11. The method of claim 7 wherein the detection method is by means of a measurement of a fluorescence signal based on amplification-mediated displacement of the CpG-specific probe.

12. The method of claim 7 wherein the amplification and detection method comprises fluorescence-based quantitative PCR.

13. The method of claim 7 wherein methylation amounts in the nucleic acid sample are quantitatively determined based on reference to a control reaction for amount of input nucleic acid.

14. A method for detecting a methylated CpG-containing nucleic acid comprising:

(a) contacting a nucleic acid-containing sample with a modifying agent that modifies unmethylated cytosine to produce a converted nucleic acid;

(b) amplifying the converted nucleic acid in the sample by means of oligonucleotide primers and in the presence of a CpG-specific oligonucleotide probe, wherein both the primers and the CpG-specific probe distinguish between modified unmethylated and methylated nucleic acid; and

(c) detecting the methylated nucleic acid based on amplification-mediated displacement of the CpG-specific probe.

15. The method of claim 14 wherein the amplifying step comprises a polymerase chain reaction (PCR).

16. The method of claim 14 wherein the modifying agent is bisulfite.

17. The method of claim 14 wherein the converted nucleic acid contains uracil in place of unmethylated cytosine residues present in the unmodified nucleic acid-containing sample.

18. The method of claim 14 wherein the detection method comprises measuring a fluorescence signal based on amplification-mediated displacement of the CpG-specific probe.

19 The method of claim 14 wherein the amplification and detection method is fluorescence-based quantitative PCR.

20. A methylation detection kit useful for the detection of a methylated CpG-containing nucleic acid comprising a carrier means being compartmentalized to receive in close confinement therein one or more containers comprising:

(i) a first container containing a modifying agent that modifies unmethylated cytosine to produce a converted nucleic acid;

(ii) a second container containing primers for amplification of the converted nucleic acid;

(iii) a third container containing primers for the amplification of control unmodified nucleic acid; and

(iv) a fourth container containing a specific oligonucleotide probe the detection of which is based on amplification-mediated displacement,

wherein the primers and probe each may or may not distinguish between unmethylated and methylated nucleic acid.

21. The kit of claim 20, wherein the modifying agent is bisulfite.

22. The kit of claim 20 wherein the modifying agent converts cytosine residues to uracil residues.

23. The kit of claim 20, wherein the specific oligonucleotide probe is a CpG-specific oligonucleotide probe, and wherein the probe, but not the primers for amplification of the converted nucleic acid, distinguishes between modified unmethylated and methylated nucleic acid.

24. The kit of claim 20, wherein the specific oligonucleotide probe is a CpG-specific oligonucleotide probe, and wherein both the probe and the primers for amplification of the converted nucleic acid, distinguish between modified unmethylated and methylated nucleic acid.

25. The kit of claim 20, wherein the probe further comprises a fluorescent moiety linked to an oligonucleotide base directly or through a linker moiety.

26. The kit of claim 20, wherein the probe is a specific, dual-labeled TaqMan® probe.

* * * * *

US005602000A

United States Patent [19]

Hyman

[11] Patent Number: 5,602,000

[45] Date of Patent: *Feb. 11, 1997

[54] **METHOD FOR ENZYMATIC SYNTHESIS OF OLIGONUCLEOTIDES**

[76] Inventor: **Edward D. Hyman**, 2100 Sawmill Rd., River Ridge, La. 70123

[*] Notice: The portion of the term of this patent subsequent to Dec. 23, 2012, has been disclaimed.

[21] Appl. No.: **464,778**

[22] Filed: **Jun. 23, 1995**

Related U.S. Application Data

[63] Continuation-in-part of Ser. No. 161,224, Dec. 2, 1993, Pat. No. 5,516,664, Ser. No. 100,671, Jul. 30, 1993, and Ser. No. 995,791, Dec. 23, 1992, Pat. No. 5,436,143.

[51] **Int. Cl.**[6] **C12P 19/34**; C12Q 1/68; C12Q 1/70; A61K 38/43

[52] **U.S. Cl.** **435/91.1**; 435/6; 435/5; 435/91.2; 424/94.1

[58] **Field of Search** 435/6, 91.1, 5; 424/94.1

[56] **References Cited**

U.S. PATENT DOCUMENTS

3,850,749	11/1974	Kaufmann et al. .	
4,385,112	8/1981	Misaki et al.	435/6
4,661,450	4/1987	Kempe et al. .	
4,987,071	1/1991	Cech et al. .	
5,256,555	10/1993	Milburn et al.	435/195
5,273,879	12/1993	Goodman et al.	435/6
5,409,817	4/1995	Tabor et al.	435/74
5,436,143	7/1995	Hyman	435/91.2

FOREIGN PATENT DOCUMENTS

0196101	10/1986	European Pat. Off. .
2169605	7/1986	United Kingdom .

OTHER PUBLICATIONS

Uhlenbeck, The Enzymes vol. XV 31–57, Acad Pres. 1984.

Shum NAR 5: 2297; 1978.

Middleton et al, Analytical Biochem, 144: 110–117, 1985.

Shum et al., "Simplified method for large scale enzymatic synthesis of oligoribonucleotides", Nucleic Acids Res. 5: 2297–2311 (1978).

Schott et al., "Single–step elogation of oligodeoxynulceotides using terminal deoxynucleotidyl transferase", Eur. J. Biochem. 143: 613–620 (1984).

Mackey et al., "New approach to the synthesis of polyribonulceotides of defined sequence", Nature 233: 551–553 (1971).

Hinton et al., "The preparative synthesis of oligodeoxyribonucleotides using RNA ligase", Nucleic Acids Res. 10: 1877–1894 (1982).

England et al., "Dinucleotide pyrophosphates are substrates for T4–induced RNA ligase", Proc. Nat'l Acad Sci. (USA) 74: 4839–4842 (1977).

Beckett et al., "Enzymatic Synthesis of Oligoribonucleotides", in *Oligonucleotide Synthesis: A Practical Approach*, M. J. Gait ed., pp. 185–197 (1984).

Mudrakovskaya et al., "RNA Ligase of Bacteriophage T4. VII: A solid pahse enymatic synthesis of oligoribonucelotides", Biorg. Khim., 17: 819–822 (1991).

Stuart et al., "Synthesis and Properties of Oligodeoxynucleotides with an AP site at a preselected location", Nulceic Acids Res. 15: 7451–7462 (1987).

Norton et al., "A ribonuclease specific for 2'–O–Methyltaed Ribonulceic Acid", J. Biol. Chem. 242: 2029–2034 (1967).

Eckstein et al., "Phosphorothioates in molecular biology", TIBS 14:97–100 (1989).

Bryant et al., "Phosphorothioate Substrates for T4 RNA Ligase", Biochemistry 21: 5877–5885 (1982).

McLaughlin et al., "Donor Activation in the T4 RNA Ligase Reaction", Biochemistry 24: 267–273 (1985).

Ohtsuka et al., "A new method for 3'–labelling of polyribonucelotides by phosphorylation with RNA ligase and its aplication to the 3'–modification for joining reactions", Nulceic Acids Res. 6: 443–454 (1979).

Kornberg, A., "Reversible Enzymatic Sysnthesis of Diphosphopyridine nucleotide and inorganic pyrophosphate", J Biol. Chem. 182: 779–793 (1950).

Kaplan et al., "Enzymatic Deamination of Adenosine Derivatives", J. Biol. Chem. 194: 579–591 (1952).

Bartkiewicz et al., "Nucleotide pyrophosphatase from potato tubers", Eur. J. Biochem. 143: 419–426 (1984).

Rand et al., "Sequence and cloning of bacteriophage T4 gene 63 encoding RNA ligase and tail fibre attachment activities", The EMBO Journal 3: 397–402 (1984).

Heaphy et al., "Effect of Single Amino Acid Chnages in the Region of the Adenylation Site of T4 RNA llgasc", Biochemistry 26: 1688–1696 (1987).

Lowe et al., "Molecular cloning and expression of a cDNA encoding the membrane–associated rat intestinal alkaline phosphatase", Biochem. Biophys. Acta 1037: 170–177 (1990).

(List continued on next page.)

Primary Examiner—W. Gary Jones
Assistant Examiner—Dianne Rees
Attorney, Agent, or Firm—Oppedahl & Larson

[57] **ABSTRACT**

Enzymatic synthesis of oligonucleotides is performed by the steps of: (a) combining a primer and a blocked nucleotide in the presence of a chain extending enzyme to form a primer-blocked nucleotide product containing the blocked nucleotide coupled to the primer at its 3'-end; (b) removing the blocking group from the 3' end of the primer-blocked nucleotide product; and (c) repeating the cycle of steps (a) and (b), using the primer-nucleotide product of step (b) as the primer for step (a) in the next cycle, for sufficient cycles to form the oligonucleotide product. Cycles may optionally include the step of converting any unreacted blocked nucleotide to an unreactive form which is substantially less active as a substrate for the chain extending enzyme. Cycles may also include the step of removing the blocking group from unreacted blocked nucleotide. This step is unnecessary, however, when the same nucleotide is added in two or more successive cycles. The synthetic cycles are preferably performed in a single vessel without intermediate purification of oligonucleotide product.

29 Claims, 10 Drawing Sheets

OTHER PUBLICATIONS

Chang et al., "Molecular Biology of Terminal Transferase", CRC Crit. Rev. Biochem. 21: 27–52.

Razzell et al., "Stduies on Polynucleotides: III. Enzymatic Degaradtion. Substrate Specificity and Properties of Snake Venom Phosphodiesterase", J. Biol. Chem. 234: 2105–2113 (1959).

Tessier et al., "Ligation of Single–Stranded Oligodeoxyribonucleotides by T4 RNA Ligase", Analytical Biochemistry 158: 171–178 (1986).

England et al., "Enzymatic Oligoribonucleotide Synthesis with T4 RNA Ligase", Biochemistry 17: 2069–2076 (1978).

Middleton et al., "Synthesis and Purificationof Oligonucleotides Using T4 RNA Ligase and Reverse–Phase Chromatography", Analytical Biochemistry 144: 110–117 (1985).

Uhlenbeck et al., "T4 RNA Ligase", The Enzymes XV: 31–58 (1982).

Hoffman et al. "Synthesis and reactivity of intermediates formed in the T4 RNA ligase reaction", Nucleic Acids Res. 15: 5289–5301 (1987).

Soltis et al., "Independent Locations of Kinase and 3'–Phosphatase Activities on T4 Polynucleotide Kinase", J. Biol. Chem. 257: 11340–11345 (1982).

Apostol et al., "Deletion Analysis of a Multifunctional Yeast tRNA Ligase Polypeptide", J. Biol. Chem. 266: 7445–7455 (1991).

Becker et al., "The Enzymatic Cleavage of Phospahte Termini from Polynucleotides", J. Biol. Chem. 242: 936–950 (1967).

Greer et al., "RNA Ligase in Bacteria: Formation of a 2',5' Linkage by an *E. coli* Extract", Cell 33: 899–906 (1983).

Schwartz et al., "Enzymatic Mechanism of an RNA Ligase from Wheat Germ", J. Biol. Chem. 258: 8374–8383 (1983).

Beabealashvilli et al., "Nucleoside 5'–triphosphates modified at sugar residues as substrats for calf thymus terminal deoxynucleotidyl transferase and for AMV reverse transcriptase", Biochim. Biophys. Acta 868: 135–144 (1986).

Lehman et al., "The Deoxyribonucleases of *Escherichia coli*", J. Biol. Chem. 239: 2628–2636 (1964).

Singer, M., "Phosphorolysis of Oligonucleotides by Polynucleotide Phosphorylase".

Itakura et al., "Syntethis and Use of Synthetic Oligonucleotides", Ann. Rev. Biochem. 33: 323–356.

Lewin (1987) "Genes: 3rd Ed." pp. 60–63 John Wiley & Sons, N.Y.

Comeron et al, Biochem 16 (23): 5120–5126 (1977).

P. T. Gilham et al, Nature, 233, 551–3, (1971).

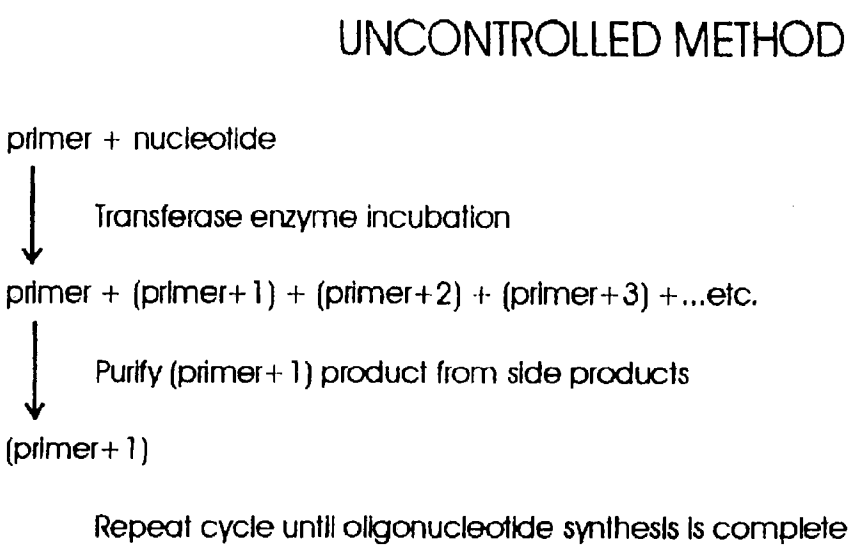

UNCONTROLLED METHOD
primer + nucleotide
Transferase enzyme incubation
primer + (primer+1) + (primer+2) + (primer+3) +...etc.
Purify (primer+1) product from side products
(primer+1)
Repeat cycle until oligonucleotide synthesis is complete

Fig. 1A

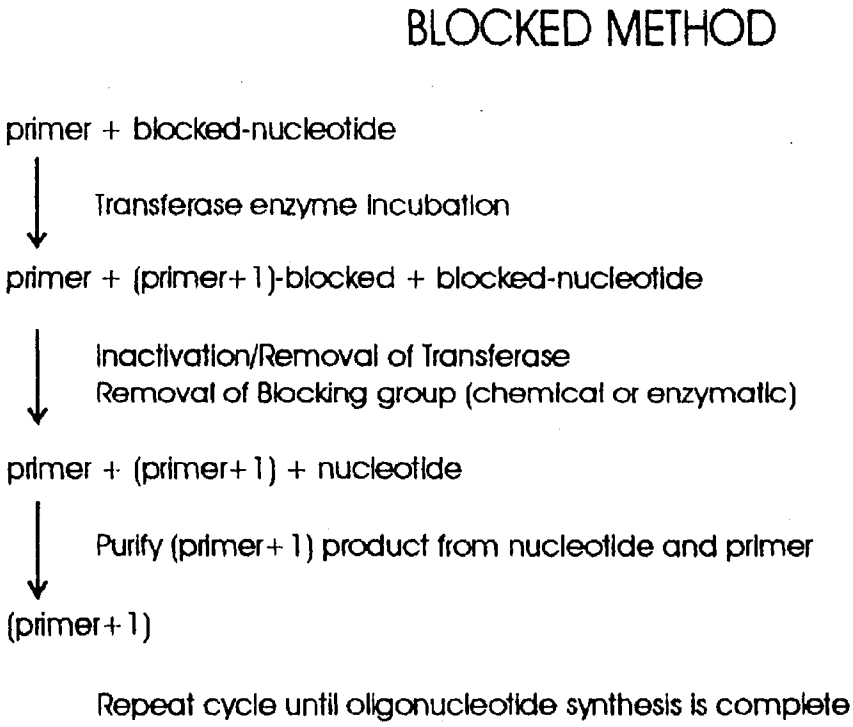

BLOCKED METHOD
primer + blocked-nucleotide
Transferase enzyme incubation
primer + (primer+1)-blocked + blocked-nucleotide
Inactivation/Removal of Transferase
Removal of Blocking group (chemical or enzymatic)
primer + (primer+1) + nucleotide
Purify (primer+1) product from nucleotide and primer
(primer+1)
Repeat cycle until oligonucleotide synthesis is complete

Fig. 1B

(A) BASIC MODE

primer + App(d)Np (or ATP + 3',5'-(d)NDP)

↓ RNA Ligase incubation, then optionally heat inactivate

primer-p(d)Np + AMP + App(d)Np

↓ Alkaline Phosphatase incubation, then heat inactivate

primer-p(d)N + Adenosine + App(d)N + PO_4

Repeat cycle until oligonucleotide synthesis is complete

Fig. 2A

(B) PREFERRED MODE

primer + App(d)Np (or ATP + 3',5'-(d)NDP)

↓ RNA Ligase incubation, heat inactivation optional

primer-p(d)Np + AMP + App(d)Np

↓ Exonuclease + Nucleotide Pyrophosphatase incubation (for example, phosphodiesterase I), then heat inactivate

primer-p(d)Np + AMP + 3',5'-(d)NDP

↓ Alkaline Phosphatase incubation, then heat inactivate

primer-p(d)N + adenosine + (deoxy)nucleoside + PO_4

Repeat cycle until oligonucleotide synthesis is complete

Fig. 2B

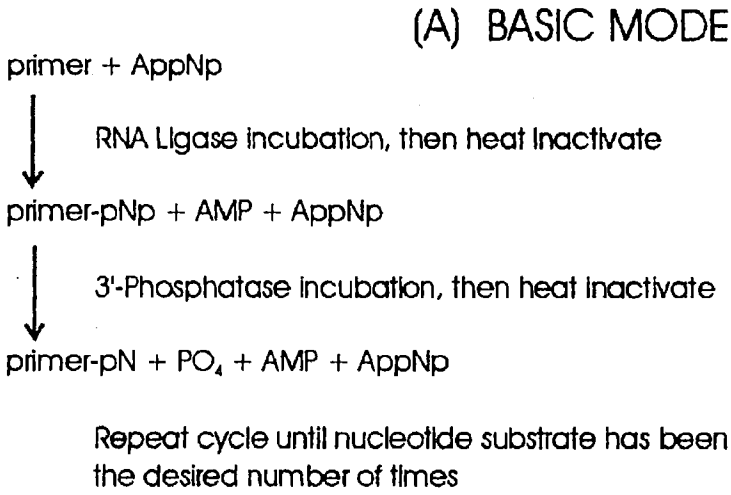

(A) BASIC MODE
primer + AppNp
RNA Ligase incubation, then heat inactivate
primer-pNp + AMP + AppNp
3'-Phosphatase incubation, then heat inactivate
primer-pN + PO_4 + AMP + AppNp
Repeat cycle until nucleotide substrate has been added to primer the desired number of times

Fig. 3A

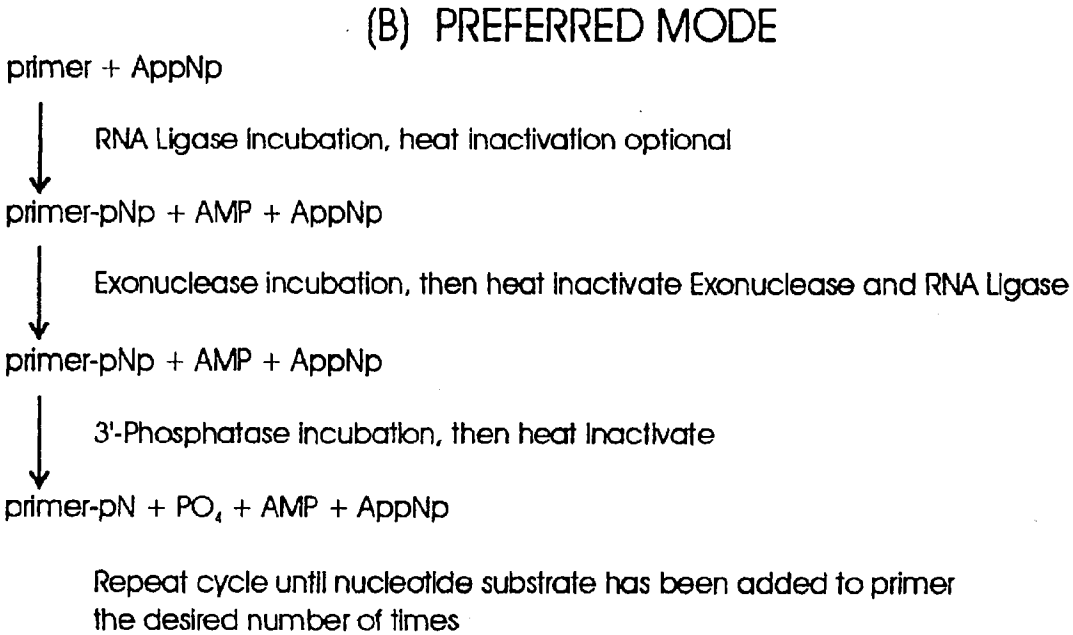

(B) PREFERRED MODE
primer + AppNp
RNA Ligase incubation, heat inactivation optional
primer-pNp + AMP + AppNp
Exonuclease incubation, then heat inactivate Exonuclease and RNA Ligase
primer-pNp + AMP + AppNp
3'-Phosphatase incubation, then heat inactivate
primer-pN + PO_4 + AMP + AppNp
Repeat cycle until nucleotide substrate has been added to primer the desired number of times

Fig. 3B

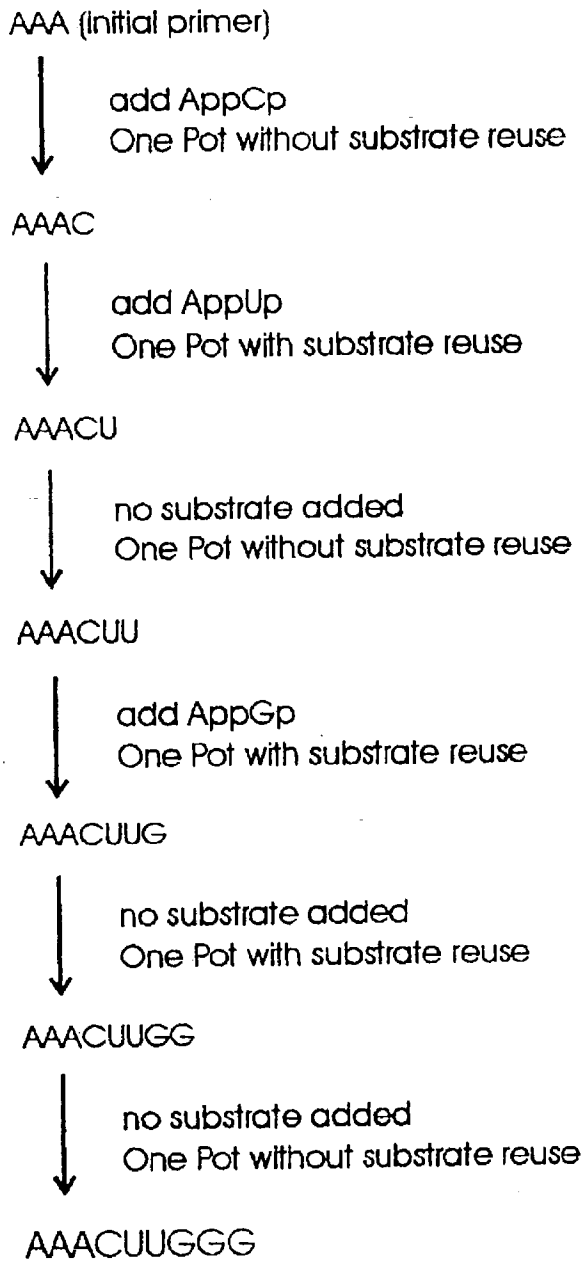
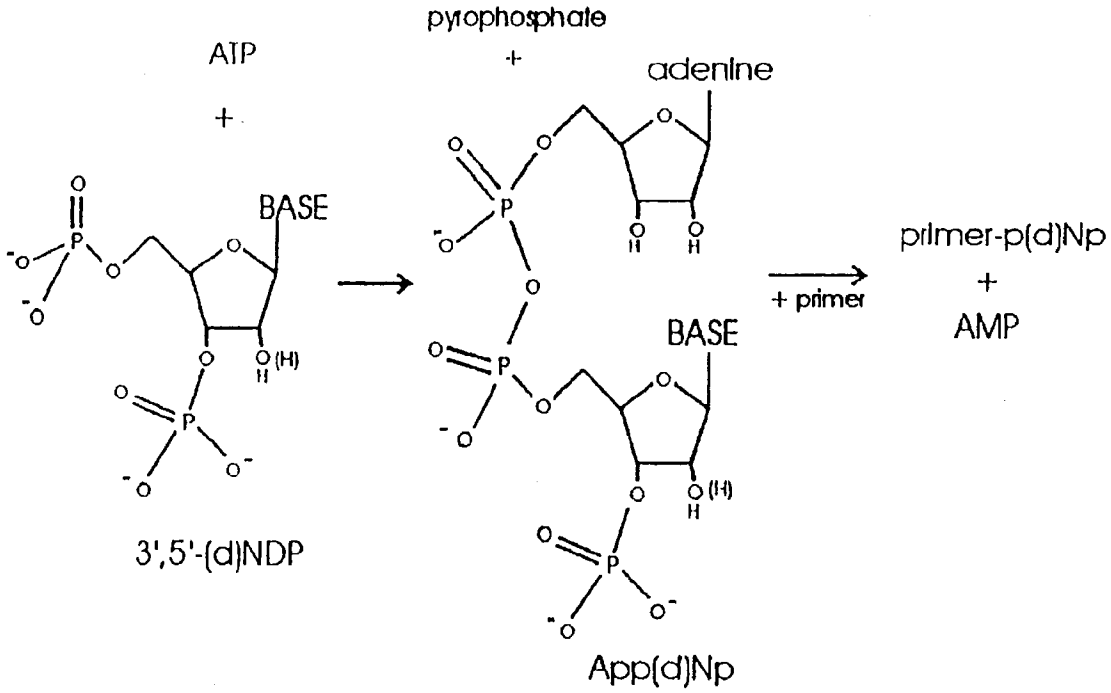

AAA (initial primer)
add AppCp
One Pot without substrate reuse
AAAC
add AppUp
One Pot with substrate reuse
AAACU
no substrate added
One Pot without substrate reuse
AAACUU
add AppGp
One Pot with substrate reuse
AAACUUG
no substrate added
One Pot with substrate reuse
AAACUUGG
no substrate added
One Pot without substrate reuse
AAACUUGGG

FIGURE 4


ATP
+
BASE
3',5'-(d)NDP
pyrophosphate
+
adenine
BASE
App(d)Np
+ primer
primer-p(d)Np
+
AMP

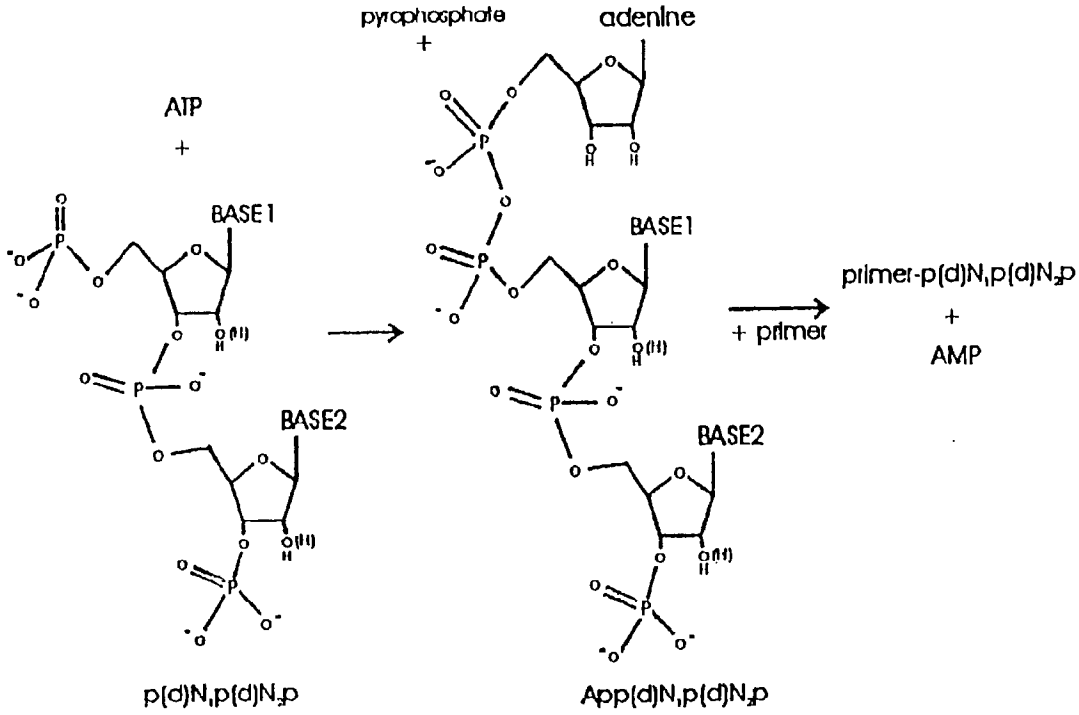

ATP
+
BASE1
BASE2
p(d)N,p(d)N'p
pyrophosphate
+
adenine
BASE1
BASE2
App(d)N,p(d)N'p
+ primer
primer-p(d)N,p(d)N'p
+
AMP

FIGURE 5

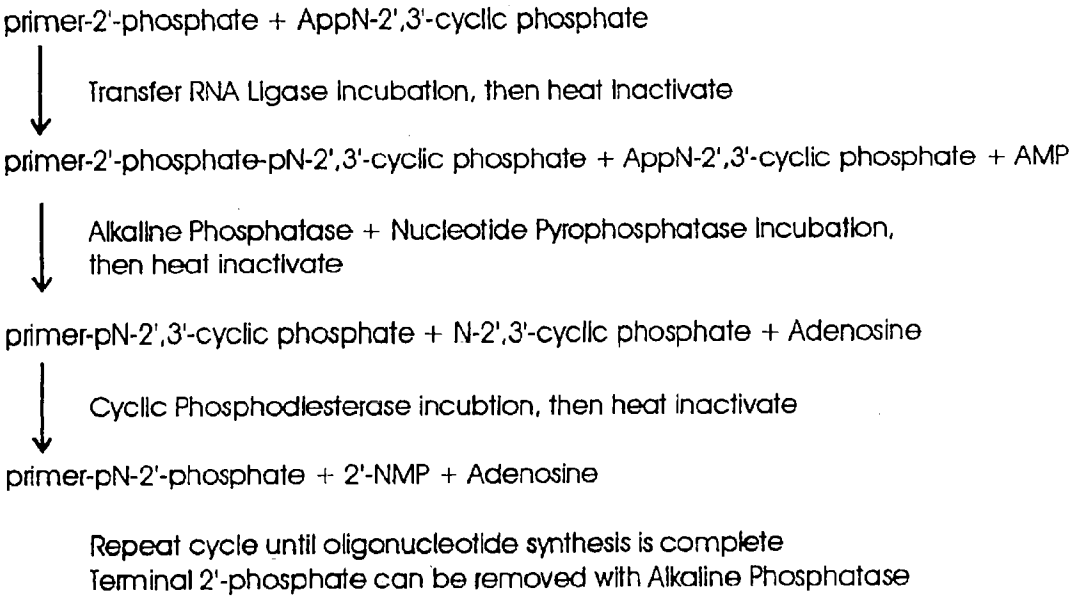

primer-2'-phosphate + AppN-2',3'-cyclic phosphate
Transfer RNA Ligase Incubation, then heat Inactivate
primer-2'-phosphate-pN-2',3'-cyclic phosphate + AppN-2',3'-cyclic phosphate + AMP
Alkaline Phosphatase + Nucleotide Pyrophosphatase Incubation, then heat inactivate
primer-pN-2',3'-cyclic phosphate + N-2',3'-cyclic phosphate + Adenosine
Cyclic Phosphodiesterase incubtion, then heat inactivate
primer-pN-2'-phosphate + 2'-NMP + Adenosine
Repeat cycle until oligonucleotide synthesis is complete
Terminal 2'-phosphate can be removed with Alkaline Phosphatase

FIGURE 6

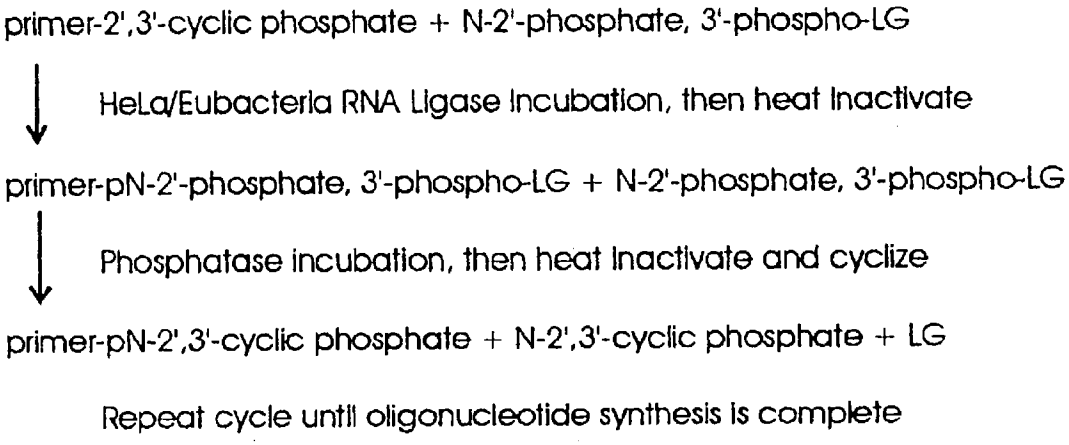

primer-2',3'-cyclic phosphate + N-2'-phosphate, 3'-phospho-LG
HeLa/Eubacteria RNA Ligase Incubation, then heat Inactivate
primer-pN-2'-phosphate, 3'-phospho-LG + N-2'-phosphate, 3'-phospho-LG
Phosphatase incubation, then heat Inactivate and cyclize
primer-pN-2',3'-cyclic phosphate + N-2',3'-cyclic phosphate + LG
Repeat cycle until oligonucleotide synthesis is complete

FIGURE 7

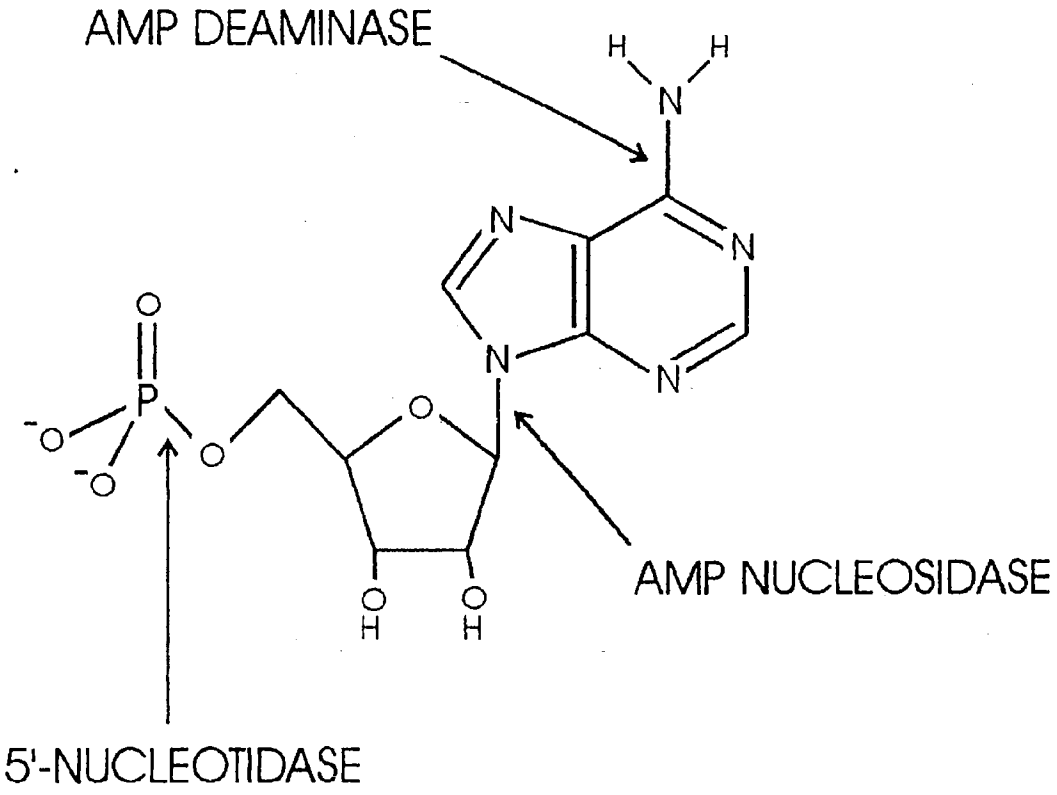


FIGURE 8

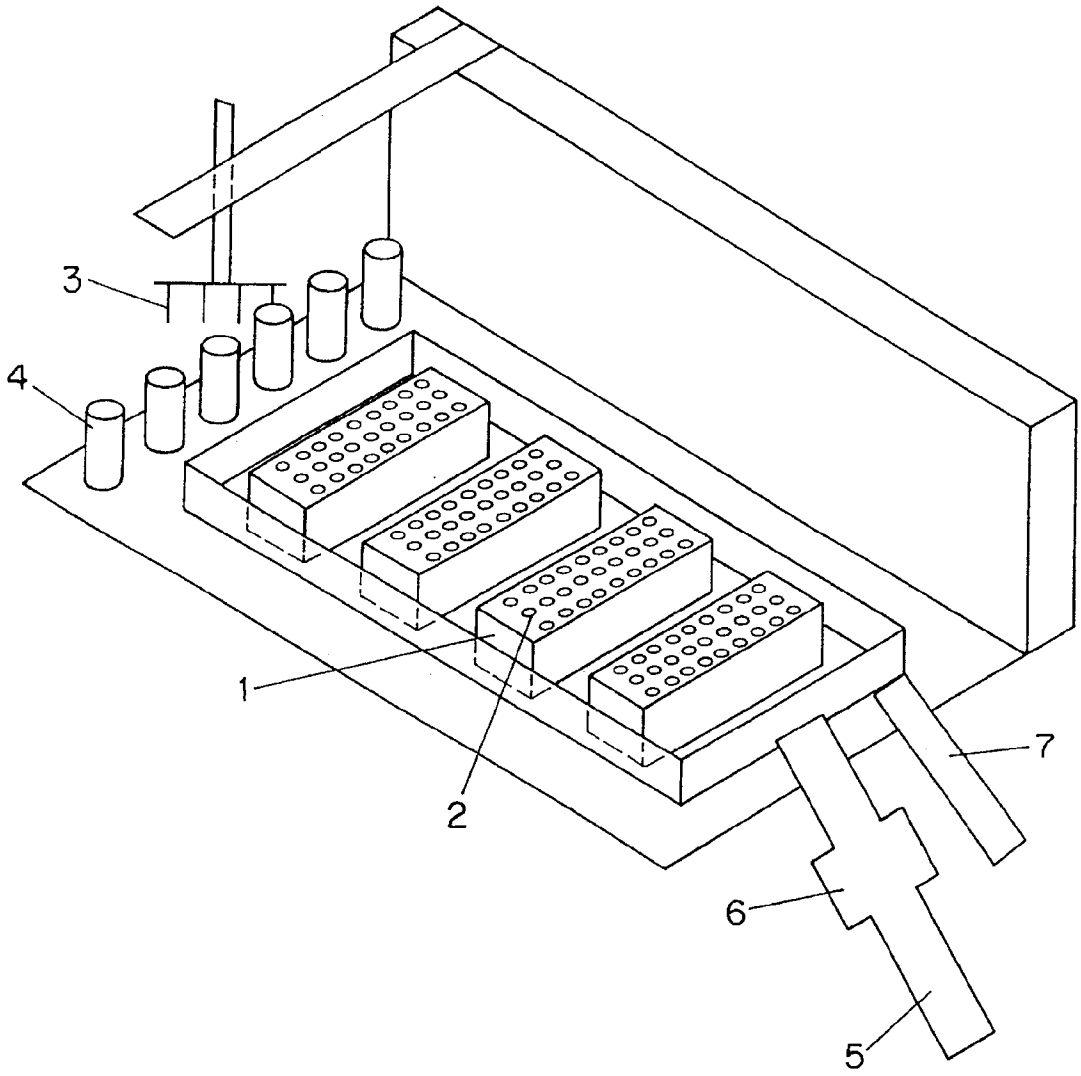


FIGURE 9

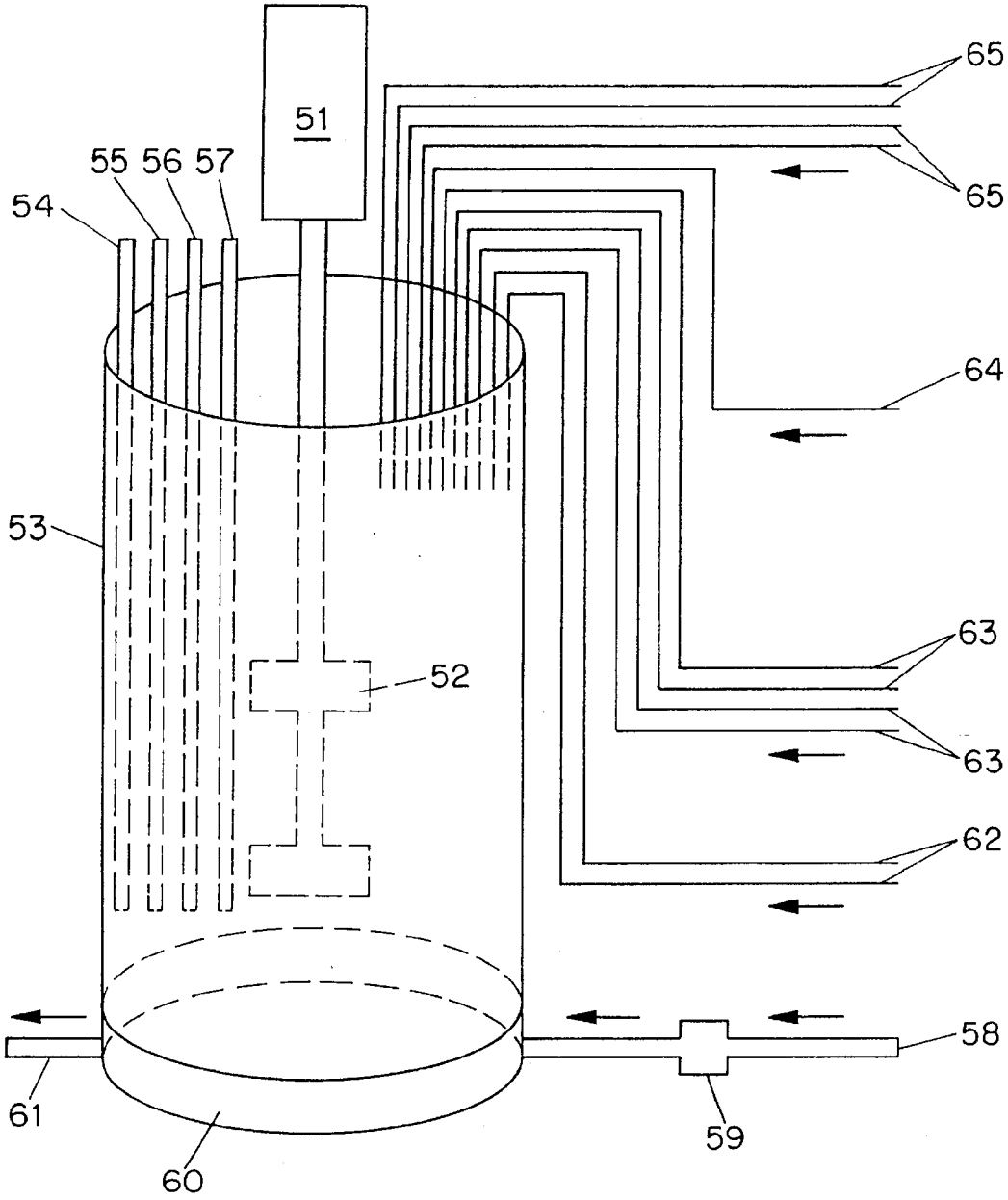


FIGURE 10

METHOD FOR ENZYMATIC SYNTHESIS OF OLIGONUCLEOTIDES

This application is a continuation-in-part of U.S. patent applications Ser. Nos. 08/161,224 filed Dec. 2, 1993, now U.S. Pat. No. 5,516,664 issued May 14, 1996; 08/100,671 filed Jul. 30, 1993 and 07/995,791 filed Dec. 23, 1992 now U.S. Pat. No. 5,436,143 issued Jul. 25, 1995.

BACKGROUND OF THE INVENTION

Synthetic oligonucleotides play a key role in molecular biology research, useful especially for DNA sequencing, DNA amplification, and hybridization. A novel "one pot" enzymatic method is described to replace both the obsolete enzymatic methods and the current phosphoramidite chemical method. This new method promises increased throughput and reliability, ease of automation, and lower cost.

Before the introduction of the phosphoramidite chemical method in 1983, enzymatic methods were used for the synthesis of oligonucleotides. Historically, two distinct enzymatic approaches have been employed as summarized in FIG. **1**. These enzymatic methods have been abandoned, however, in favor of the superior phosphoramidite chemical method.

The first enzymatic approach is the "uncontrolled" method. As depicted in FIG. **1**A, a short oligonucleotide primer is incubated with the desired nucleotide and a nucleotidyl transferase. At the end of the optimal incubation period, a mixture of oligonucleotide products containing different numbers of bases added to the primer (i.e. primer, primer +1,primer +2 . . .) is obtained. The desired product, the primer with one added base, is purified using either electrophoresis or chromatography. The process of enzyme incubation and oligonucleotide purification is repeated until the desired oligonucleotide is synthesized. Examples of the use of this approach are: (1) Polynucleotide Phosphorylase ("PNP") and ADP, GDP, CDP, and UDP have been used to make oligodeoxyribonucleotides in accordance with the following reaction:

$$\text{primer}+\text{rNDP}\rightarrow(\text{primer-rN})\text{-3'-OH}+PO_4$$

Shum et al, *Nucleic Acids Res.*, 5(7): 2297-311 (1978), and (2) Terminal deoxynucleotidyl Transferase and the nucleotides dATP, dGTP, dCTP, and dTTP have been used to make oligodeoxyribonucleotides in accordance with the following reaction:

$$\text{primer}+\text{dNTP}\rightarrow(\text{primer-dN})\text{-3'-OH}+\text{pyrophosphate}$$

Schott et al, *Eur. J. Biochem,* 143: 613–20 (1984). The flaws of the "uncontrolled" approach are the requirement for cumbersome manual purification of the primer+**1** product after each coupling cycle, poor yields of the desired primer+**1** product, and inability to automate.

The second enzymatic approach is the "blocked" method, shown in FIG. **1**B. The nucleotide used in the extension step is blocked in some manner to prevent the nucleotidyl transferase from adding additional nucleotides to the oligonucleotide primer. After the extension step, the oligonucleotide product is separated from the enzyme and nucleotide, and the blocking group is removed by altering the chemical conditions or by the use of a second enzyme. The oligonucleotide product is now ready for the next extension reaction. Examples of this approach are: (1) PNP and NDP-2'-acetal blocked nucleotides have been used to make oligoribonucleotides. The acetal blocking group is removed under acidic conditions (Gilham et al, *Nature,* 233: 551–3 (1971) and U.S. Pat. No. 3,850,749), (2) RNA ligase and the blocked nucleotide App(d)Np (or ATP+3',5'-(d)NDP) have been used to make oligoribonucleotides and oligodeoxyribonucleotides. The 3'-phosphate blocking group is removed enzymatically with a phosphatase such as alkaline phosphatase (T. E. England et al, *Biochemistry,* (1978), 17(11), 2069–81; D. M. Hinton et al, *Nucleic Acids Research,* (1982), 10(6), 1877–94).

The advantage of the "blocked" method over the "uncontrolled" method is that only one nucleotide can be added to the primer. Unfortunately, the "blocked" method has several flaws which led to its abandonment in favor of the chemical method. The "blocked method", like the "uncontrolled" method, requires the purification of the oligonucleotide product from the reaction components after each coupling cycle.

In the first approach, using PNP, the oligonucleotide is exposed to acid to remove the acid-labile acetal blocking group. Oligonucleotide product must be purified and redissolved in fresh buffer in preparation for the next polymerization reaction for two reasons: (1) PNP requires near neutral pH conditions whereas acetal removal requires approximately pH 1; and (2) the product of the polymerization reaction, PO_4, must be removed or it will cause phosphorolysis of the oligoribonucleotide catalyzed by PNP.

In the second approach, using RNA ligase, the art teaches that oligonucleotide product needs to be purified after each cycle because the dinucleotide App(d)N, formed by phosphatase treatment of App(d)Np, is still a suitable substrate for RNA ligase and must be completely removed prior to addition of RNA ligase in the next cycle. England et al., *Proc. Natl. Acad. Sci. USA,* 74(11): 4839–42 (1977). Hinton et al. emphasize the importance of purifying oligonucleotide product after each cycle by stating: "This elution profile [a DEAE-sephadex chromatogram of oligodeoxyribonucleotide product] also demonstrates the absence of either significant contaminating products arising from nucleases or of the reaction intermediate, A-**5'**pp**5'**-dUp. The absence of such substances is critical if this general methodology is to be useful for synthesis." Hinton et al, *Nucleic Acids Research,* 10(6):1877–94 (1982). The art also teaches that nucleoside and phosphate by-products generated by phosphatase incubation of the RNA Ligase reaction mixture substantially inhibit RNA Ligase activity and must be removed prior to subsequent RNA ligation steps in order to work usefully. Middleton et al., *Anal. Biochem.,* 144:110–117 (1985).

Two modifications have been devised for the "blocked" method to improve the oligonucleotide product yield and to speed required oligonucleotide product purification after each coupling cycle. The first modification was .the use of a branched synthetic approach (*Oligonucleotide Synthesis: a practical approach,* M. J. Gait editor, (1985), pp. 185–97, IRL Press). This approach improved the yield of final oligonucleotide product, but intermediate purification of oligonucleotide after each coupling cycle was still required. The second modification was the covalent attachment of the primer chain to a solid phase support (A. V. Mudrakovskaia et al, *Bioorg. Khim,* (1991), 17(6), 819–22). This allows the oligonucleotide to be purified from all reaction components simply by washing the solid phase support column. However, product yields are still low, and primer chains which do not couple during a cycle are not removed and are carried over to the next coupling cycle. It appears that the poor coupling efficiency results from steric problems encountered

by the enzyme in gaining access to the covalently bound primer chain. Unfortunately, it is not possible to combine these two modifications in an automated manner. The current phosphoramidite chemical method for oligonucleotide synthesis also utilizes a solid phase support to facilitate oligonucleotide purification after each coupling reaction.

The present invention provides a method for enzymatic oligonucleotide synthesis which is preferably performed entirely in a single tube, requiring only temperature control and liquid additions, and not requiring intermediate purifications or solid phase supports. This method is well suited for automation on a liquid handling robot apparatus, allowing the simultaneous preparation of a thousand oligonucleotides per day in microtiter plates. This capability dwarfs the best commercially available instrument which can prepare only four oligonucleotides simultaneously with the phosphoramidite method (Applied Biosystems, Inc.).

SUMMARY OF THE INVENTION

This invention provides a method for enzymatic synthesis of oligonucleotides of defined sequence. The method involves the steps of:

(a) combining an oligonucleotide primer and a blocked nucleotide, or a blocked nucleotide precursor that forms a blocked nucleotide in situ in a reaction mixture, in the presence of a chain extending enzyme effective to couple the blocked nucleotide to the 3'-end of the oligonucleotide primer such that a primer-blocked nucleotide product is formed, wherein the blocked nucleotide comprises a nucleotide to be added to form part of the defined sequence and a 3'- blocking group attached to the nucleotide effective to prevent the addition of more than one blocked nucleotide to the primer;

(b) removing the blocking group from the 3'-end of the primer-blocked nucleotide product to form a primer-nucleotide product; and

(c) repeating at least one cycle of steps (a) and (b) using the primer-nucleotide product from step (b) as the oligonucleotide primer of step (a) of the next cycle, without prior separation of the primer-nucleotide product from the reaction mixture, using blocked nucleotides appropriate to the defined sequence of the oligonucleotide being synthesized.

When the defined sequence calls for the same nucleotide to be incorporated more than once in succession, unreacted blocked nucleotide may be reused in the subsequent cycle(s). In this case, the blocking group is selectively removed from the primer-blocked nucleotide product substantially without deblocking of the unreacted blocked nucleotide. Otherwise, the method includes the further step of converting any unreacted blocked nucleotide to an unreactive form which is substantially less active as a substrate for the chain extending enzyme than the blocked nucleotide. The method of the invention is preferably performed in a single reaction vessel, without intermediate purification of oligonucleotide product.

In accordance with one embodiment of the invention, a single cycle comprises the steps in sequence:

(a) incubation of an oligonucleotide primer with RNA ligase and App(d)Np or App(d)N_1p(d)N_2p or precursors thereof, wherein App is an adenosine diphosphate moiety, and Np, N_1 and N_2 are a 3'-phosphate-blocked nucleoside moiety, to form a primer-pNp product;

(b) incubation with a Phosphatase; and

(c) heat inactivation of the Phosphatase.

By careful selection of the conditions of the reaction with the Phosphatase, the selectivity of the enzymatic dephosphorylation reaction can be controlled, such that unreacted blocked nucleotide substrate is either substantially inactivated when it is not to be reused, and substantially left intact when reuse is desired.

In accordance with a preferred embodiment, a single cycle of the method comprises the steps in sequence:

(a) incubation of an oligonucleotide primer with RNA ligase and App(d)Np or App(d)N_1p(d)N_2p or precursors thereof;

(b) incubation with an exonuclease and a nucleotide pyrophosphatase (e.g. snake venom phosphodiesterase I);

(c) heat inactivation of the Exonuclease and Nucleotide Pyrophosphatase;

(d) incubation with a Phosphatase; and

(e) heat inactivation of the Phosphatase.

BRIEF DESCRIPTION OF THE DRAWINGS

FIGS. **1**A and B: The "Uncontrolled" and "Blocked" enzymatic methods previously used for the synthesis of oligonucleotides.

FIGS. **2**A and B: The method of the invention for the synthesis of oligonucleotides.

FIGS. **3**A and B: The method of the invention for the synthesis of repeat regions of an oligonucleotide.

FIG. **4**: Synthesis of repeat and non-repeat regions using the method of the invention

FIG. **5**: Reactions catalyzed by RNA ligase.

FIG. **6**: An embodiment of the invention utilizing Transfer RNA ligase-as the chain extending enzyme.

FIG. **7**: An embodiment of the invention utilizing HeLa/Eubacterial RNA Ligase as the chain extending enzyme.

FIG. **8**: The structure of AMP and the cleavage points of various enzymes.

FIG. **9**: Apparatus for practicing the method of the invention, suitable for synthesizing many oligonucleotides simultaneously.

FIG. **10**: Apparatus for practicing the method of the invention, suitable for the bulk synthesis of an oligonucleotide.

DETAILED DESCRIPTION OF THE INVENTION

The present invention provides a method for synthesizing oligonucleotides enzymatically which can be performed in a single vessel without the need for any intermediate purification step. An embodiment of the method of the invention in Basic Mode is shown in FIG. **2**A. In this embodiment, new nucleotide substrate is added for each new cycle.

As shown in FIG. **2**A, a reaction mixture is formed containing an oligonucleotide primer, a blocked nucleotide substrate, and a chain extending enzyme such as RNA ligase and is incubated to couple the blocked nucleotide to the oligonucleotide primer. The RNA ligase may then be inactivated, for example by heating. The resulting reaction mixture contains the primer-blocked nucleotide product, unreacted primer, unreacted blocked nucleotide, and adenosine monophosphate (AMP). Alternatively, the RNA ligase may be left in active form and the substrate rendered inactive for further reaction with the primer.

The next step as shown in FIG. **2**A is incubation with an enzyme which removes the blocking group from the primer-blocked nucleotide product and unreacted blocked nucle-

otide. The resulting reaction mixture, containing unreacted primer, extended primer, and unblocked nucleotide substrate can then be recycled directly for use as the primer in the subsequent cycle without performing intermediate purification of extended primer. Such intermediate purification is taught by prior art as an essential step.

FIG. 3A shows an alternative embodiment of the invention, in which the unreacted blocked nucleotide is recycled to form a region of the oligonucleotide in which the same base is repeated. For example, the 8-mer oligonucleotide 5'-AGUGGCCC-3' contains a consecutive repeat of G and two consecutive repeats of C. Synthesizing the repeat region of this oligonucleotide using the method shown in FIG. 2A results in a significant waste of materials. In this situation it may be preferable when synthesizing the oligonucleotide not to inactivate or deblock the unreacted nucleotide substrate during a cycle, so that the unreacted nucleotide can be reused in the ensuing cycle. This is accomplished by a modification of the method of FIG. 2A which is outlined in FIG. 3A.

As shown in FIG. 3A, the first step is again the addition of a blocked nucleotide to the 3'-end of the primer. In this case, however, the blocking group is selectively removed from the primer-blocked nucleotide product without significantly deblocking, and thus inactivating, the unreacted nucleotide in the reaction mixture using a 3'-phosphatase. The unblocked primer-nucleotide product is then used as the primer for the next cycle and unreacted blocked nucleotide is used as the blocked nucleotide of the next cycle. Similar to the method of FIG. 2A, the modified method for synthesis of repeat regions may be performed without intermediate purification of the extended primer product. This method may be employed for as many cycles as necessary until the repeat region is synthesized.

In the synthesis of an oligonucleotide with at least one repeat region and at least one non-repeat region, cycles of both methods shown in FIGS. 2A and 3A may be employed to provide an overall synthetic strategy in which repeat regions are synthesized using either method, but preferably the method of FIG. 3A, and non-repeat regions are synthesized using the method of FIG. 2A. A hypothetical synthesis is shown in FIG. 4.

The method of the invention is surprisingly useful because problems identified in the prior art which suggest that the method would not work, have been found by the inventor not to limit the utility of the invention. Prior art (Hinton et al.) teaches that the extended primer product must be separated from the reaction mixture to remove App(d)N, which is able to couple to the primer in the next cycle. It is the discovery of the inventor that the unblocked nucleotide, App(d)N, is substantially less active as a substrate for RNA ligase (e.g. 50 to 100 times less active) than the blocked nucleotide, App(d)Np, obviating the need for separating the unblocked nucleotide from the extended primer product. Prior art (Middleton et al.) also teaches that the nucleoside and phosphate by-product of the phosphatase incubation substantially inhibit RNA ligase, and must be separated from the extended primer product at the end of each cycle in order to work usefully. It is the discovery of the inventor that the by-products of the enzymatic reactions do not significantly inhibit the enzymes, especially RNA ligase.

Experiments were performed by the inventor on each of the reaction by-products confirm the absence of significant inhibition of RNA ligase. The major by-products of the method of the invention are nucleotides and PO_4. No inhibition was detected in the presence of 10 mM PO_4 and 10 mM Adenosine (a typical nucleoside). Extremely weak inhibition was observed in the presence of 100 mM PO_4. In addition, other nucleotides were tested for inhibition: no inhibition was detected in the presence of 10 mM Adenine, 1 mM AMP, 1 mM ATP, 2 mM AppA and 10 mM 3',5'-ADP; extremely weak inhibition was detected in the presence of 10 mM Pyrophosphate; and strong inhibition was observed in the presence of 10 mM AMP and 10 mM ATP. Therefore, the only two products which are strong inhibitors, ATP and AMP, and one product which is an extremely weak inhibitor, pyrophosphate, will never accumulate to these high concentrations since they are degraded by Alkaline Phosphatase.

After the completion of the appropriate number of cycles, the synthesized oligonucleotide may be used in some applications without purification. Alternatively, if purification is required, this can be accomplished using known methods: centrifugation, extraction with organic solvents such as phenol, chloroform and ethyl ether; precipitation, e.g. using ethanol or isopropanol in the presence of high salt concentration; size exclusion, anion exchange, reverse phase, or thin layer chromatography; ultrafiltration or dialysis; gel electrophoresis; hybridization to a complementary oligonucleotide; or by an affinity ligand interaction, such as biotin-avidin. The oligonucleotide may also be attached to a solid support throughout its synthesis, e.g., via the primer, in which case final purification may be performed by washing the support.

The method of the invention may also be used in combination with other methods for synthesizing oligonucleotides such that the method of the invention is used to make a portion of the final oligonucleotide product. Such other methods may include the blocked enzymatic method, the uncontrolled enzymatic method, the branched enzymatic method, chemical methods, transcription-based enzymatic methods, template-based enzymatic methods, and post-synthetic modification methods.

The method of the invention offers numerous advantages by operating in a mild aqueous system. The specificity of the enzymatic reactions obviates the need for base protecting groups, highly reactive functional groups, and harsh solvent conditions. All nucleotide and enzyme reagents are non-hazardous and are stable at room temperature in aqueous solution. In contrast, the phosphoramidite chemical method is encumbered with hazardous solvents, unstable nucleotides, harsh acids and bases, and solid phase supports.

PRIMERS

The primer used in the first cycle of the method of the invention, denoted as the "initial primer" herein, is an oligonucleotide of length sufficient to be extended by the chain extending enzyme. For example, if RNA ligase is the chain extending enzyme, the length is usually at least three-bases. Primers for use in the invention can be made using known chemical methods, including the phosphoramidite method. Other methods include DNase or RNase degradation of synthetic or naturally occurring DNA or RNA. Numerous primers suitable for use in the invention are commercially available from a variety of sources. The initial primer may be selected to provide the first three bases of the ultimate product, or it may be selected to provide facile cleavage of some or all of the initial primer to yield the desired ultimate product.

In most applications, the presence in the oligonucleotide product of the 5'-extension corresponding to the initial primer is inconsequential. These applications may include

DNA sequencing, polymerase chain reaction, and hybridization. However, some applications may necessitate the removal of all or part of this 5'-extension. Several procedures have been designed to achieve this result. These procedures are based on a structural or sequence difference between the initial primer and the synthesized oligonucleotide, such that an enzyme can detect the difference and cleave the oligonucleotide into two fragments: an initial primer fragment and the desired synthesized oligonucleotide fragment. Such procedures preferably require only liquid addition to the oligonucleotide solution, and can be categorized by the type of synthesized oligonucleotide for which a procedure can be used: oligodeoxyribonucleotides, oligoribonucleotides, or both types.

OLIGODEOXYRIBONUCLEOTIDES:

(1) Initial primers containing a 3' terminal ribose can be cleaved off with either RNase or alkali. RNase, such as RNase A or RNase One (Promega), hydrolyzes only at the ribose bases of an oligonucleotide.

(2) Initial primers containing a 3' terminal deoxyuridine base can be cleaved off by incubation with Uracil DNA Glycosylase, followed by base catalyzed beta elimination. Stuart et al, *Nucleic Acids Res.*, 15(18): 7451–62 (1987).

OLIGORIBONUCLEOTIDES:

(1) Initial primers containing a 3' terminal deoxyribose base can be cleaved off with DNase. Examples of RNase-free DNases include DNase I and DNase II.

OLIGODEOXYRIBONUCLEOTIDES AND OLIGORIBONUCLEOTIDES:

(1) If the initial primer contains an appropriate recognition sequence then the initial primer can be cleaved off by incubation with an appropriate ribozyme. Alternatively, the initial primer can itself be a ribozyme containing the ribozyme recognition sequence. Cleavage is performed by adjusting reaction conditions or adding a necessary cofactor to turn on the dormant ribozyme activity.

(2) If the initial primer contains an appropriate recognition sequence, then the initial primer can be cleaved off by incubation with an appropriate single-strand-recognizing restriction endonuclease. Examples of such endonuclease include Hha I, HinP I, MnI I, Hae III, BstN I, Dde I, Hga I, Hinf I, and Taq I (New England Biolabs catalog).

(3) If the initial primer contains a 3'-terminal 2'-O-methyl ribose base, then the initial primer can be cleaved off by incubation with RNase alpha (J. Norton et al, *J. Biol. Chem.*, (1967), 242(9), 2029–34). RNase alpha cuts only at bases containing a 2'-O-methyl ribose sugar.

(4) If the initial primer is composed of some ribose bases, an oligodeoxyribonucleotide specifically annealing to the initial primer and RNase H can be added to cleave off the initial primer.

(5) If the initial primer is composed of some ribose bases, an oligoribonucleotide specifically annealing to the initial primer and a double strand specific RNase such as RNase V1 can be added to cleave the initial primer. If the initial primer is self-annealing, addition of an annealing oligoribonucleotide would not be necessary.

(6) An oligodeoxyribonucleotide may be added which anneals to the initial primer and forms a double stranded DNA region. The initial primer may then be cleaved by addition of an appropriate restriction enzyme. The initial primer can also be a self-annealing oligodeoxyribonucleotide, obviating the need to add an annealing oligodeoxyribonucleotide.

(7) If the initial primer contains a unique ribose base absent from the synthesized oligonucleotide, then the initial primer can be cleaved by incubation with an appropriate base-specific Ribonuclease. Examples include RNase CL_3 (cleaves after cytosine only), RNase T_1 (cleaves after guanosine only), and RNase U_2 (cleaves after adenosine only).

(8) If the synthesized oligonucleotide contains at least one phosphorothioate internucleotidic linkage, and the initial primer does not contain any phosphorothioate internucleotidic linkages, then the initial primer can be cleaved off by incubation with an appropriate nuclease or 5'→3' exonuclease, which is unable to hydrolyze phosphorothioate internucleotidic linkages, or hydrolyzes them poorly.

After cleaving off the initial primer from the synthesized oligonucleotide, the initial primer may be selectively degraded to nucleosides or nucleotides. This technique is based on the differential presence of a terminal phosphate monoester on the initial primer and on the synthesized oligonucleotide and the use of differential digestion with an appropriate exonuclease. Three techniques may be employed.

If the cleavage results in a 5'-phosphate on the synthesized oligonucleotide fragment and a 5'-hydroxyl on the initial primer fragment, then subsequent incubation with spleen phosphodiesterase II (a 5' to 3' exonuclease) will selectively hydrolyze the initial primer fragment to nucleotides. The 5'-phosphate protects the synthesized oligonucleotide from hydrolysis.

If the cleavage results in a 3'-hydroxyl group on the initial primer fragment and a 3'-phosphate on the synthesized oligonucleotide fragment, the initial primer fragment can be degraded using a 3' to 5' exonuclease. This can be accomplished by cleaving off the initial primer prior to the removal of the terminal 3'-phosphate blocking group from the synthesized oligonucleotide. Suitable exonucleases include exonuclease I, phosphodiesterase I and polynucleotide phosphorylase.

If the cleavage results in a 5'-hydroxyl group on the synthesized oligonucleotide fragment and a 5'-phosphate on the initial primer, then the initial primer fragment can be degraded using a 5' to 3' exonuclease with a substantial preference for 5'-phosphate substrates such as lambda exonuclease. This can be accomplished by phosphorylating the oligonucleotide at the 5'-end prior to cleavage, e.g. using polynucleotide kinase.

The cleavage of the oligonucleotide and digestion of the initial primer can be performed at any cycle of the synthesis. For bulk synthesis of a single oligonucleotide, it is preferably performed at the end of the synthesis. For synthesis of multiple oligonucleotides simultaneously, it is preferably performed after synthesizing the first three bases of the oligonucleotide. Further, it will be appreciated that the cleavage does not necessarily need to occur at the junction of the initial primer region and the synthesized oligonucleotide region.

CHAIN EXTENDING ENZYME

The chain extending enzyme used in the method of the invention is preferably RNA ligase. RNA ligase is commercially available from numerous suppliers and has been well characterized in the literature. The reactions catalyzed by RNA ligase relevant to the invention are shown in FIGS. 5A and B.

RNA ligase possesses a number of properties which make it particularly useful in the invention:

(1) The coupling reaction catalyzed by RNA ligase is thermodynamically favorable. In the presence of an

AMP inactivating enzyme, the coupling reaction is irreversible.

(2) RNA ligase couples numerous nucleotide analogs, allowing the synthesis of oligonucleotides containing these analogs using the method of the invention. Modifications include base analogs, sugar analogs, and internucleotide linkage analogs. Uhlenbeck et al, *The Enzymes,* vol. xv, pp. 31–58, Academic Press (1982) and Bryant et al, *Biochemistry,* 21:5877–85 (1982).

(3) RNA ligase couples both ribose and deoxyribose nucleotides, allowing the synthesis of oligodeoxyribonucleotides, oligoribonucleotides, and mixed ribose/deoxyribose oligonucleotides using the method of the invention.

(4) RNA ligase nucleotide substrate can be up to two bases in length in the method of the invention; i.e., App(d)N_1p(d)N_2p or p(d)N_1p(d)N_2p..

While RNA Ligase is the preferred chain extending enzyme for use in the present invention, other enzymes are within the scope of the invention. For example, because T4 RNA Ligase requires replenishment after each cycle due to its thermal instability, further refinement of the method is anticipated by the use of a thermostable RNA Ligase. A thermostable RNA Ligase is workable since the presence of RNA Ligase in other steps of a cycle is not deleterious. A thermostable RNA Ligase could be added in the first cycle and would not need replenishment throughout the oligonucleotide synthesis, reducing the expense of RNA Ligase per synthesis. Furthermore, a thermostable RNA Ligase with activity at elevated temperatures (65° to 95° C.) may provide the added benefit of reducing primer secondary structure interference with the coupling reaction. Another potential benefit of a thermostable enzyme is high activity at high ionic strength. One probable source of a thermostable RNA Ligase is thermophilic archeabacteria.

Man-made genetic mutants of T4 RNA Ligase useful in the invention without modification include a mutant version with the improved ability to extend an oligodeoxyribonucleotide primer, and a mutant version which is not inactivated at elevated temperatures.

Several other enzymes are denoted in the literature as "RNA Ligases", i.e., Transfer RNA Ligase and HeLa/Eubacterial RNA Ligase. These enzymes differ from T4 RNA Ligase in their substrate requirements in that they are reported in the literature as unable to extend a primer containing a 2'-hydroxyl, 3'-hydroxyl terminus. Consequently, they are not considered as RNA Ligase in this invention Nevertheless, these other enzymes do have the ability to act as chain extending enzymes within the scope of the present invention.

Transfer RNA Ligase is reported in the scientific literature to catalyze a reaction similar to T4 RNA Ligase, but absolutely requiring a primer with a 2'-phosphate and 3'-hydroxyl terminus. Transfer RNA Ligase has been characterized in several eukaryotes, including yeast (Apostol et al, *J. Biol. Chem.,* 266:7445–55 (1991)) and wheat germ (Schwartz et al, *J. Biol. Chem.,* 258: 8374–83 (1983)). Based on the fact that it is essential in transfer RNA processing, Transfer RNA Ligase should be ubiquitous in eukaryotes. Transfer RNA Ligase is a single polypeptide containing three distinct enzyme activities: ligase, cyclic phosphodiesterase, and 5'-polynucleotide kinase. It is the Ligase activity which catalyzes the ligation reaction described above for Transfer RNA Ligase. Since these separate activities have been mapped to separate locations on the polypeptide, it is conceivable that a mutant (e.g. a deletion mutant) can be constructed which contains only the ligase activity.

An embodiment of the method of the invention employing Transfer RNA Ligase or the mutant form as a chain extending enzyme is shown in FIG. **6**. Blocked nucleotide substrate, AppN-2',3'-cyclic phosphate, is coupled to a primer-2'-phosphate by the ligase. The second step is inactivation of unreacted blocked nucleotide substrate with Nucleotide Pyrophosphatase, e.g. snake venom phosphodiesterase I, and removal of the 2'-phosphate with a Phosphatase, e.g. Alkaline Phosphatase. (Phosphatase removal of 2'-phosphate may be unnecessary). The Phosphodiesterase I also removes unextended primer chains. The third step is incubation with cyclic phosphodiesterase to remove the blocking group from the 3' end of the extended primer by converting the terminal 2',3'-cyclic phosphate to 2'-phosphate. Such a cyclic phosphodiesterase enzyme is one of the components of Transfer RNA Ligase, whose activity has been isolated by mutation. Apostol et al., *J Biol. Chem.* 266:7445–7455 (1991). The cycle is then repeated until the desired sequence is obtained. Conceivably, the nucleotide substrate reuse technique can also be implemented if Nucleotide Pyrophosphatase is not added and the cyclic phosphodiesterase has the desired substrate selectivity.

HeLa/Eubacterial RNA Ligase catalyzes the reaction: primer-2',3'-cyclic phosphate+5'-hydroxyl-nucleotide substrate →primer-nucleotide, by direct nucleophilic attack of the 5'-hydroxyl of the nucleotide substrate on the cyclic phosphate. The HeLa RNA Ligase forms a normal 3'-5'-phosphodiester linkage; the Eubacterial RNA Ligase forms an unusual 2'-5' phosphodiester linkage (Greer et al, *Cell,* vol. 33, 899–906). An embodiment of the invention employing HeLa or Eubacterial RNA Ligase as the chain extending enzyme is shown in FIG. **7**. N-2'-phosphate, 3'-phospho-LG is used as the blocked nucleotide substrate, wherein LG is a good leaving group for nucleophilic displacement (such as dinitro-phenol or 5'-AMP) and the nucleoside N has a free 5'-hydroxyl. The first step is HeLa or Eubacterial RNA Ligase incubation with a primer-2',3'-cyclic phosphate and blocked nucleotide substrate to form primer-blocked nucleotide product. The second step is Phosphatase incubation to remove the 2'-phosphate protecting group. Spontaneously or upon heating, the terminal 3'-phospho-LG will cyclize non-enzymatically to form 2',3'-cyclic phosphate. The cyclized unreacted nucleotide is probably a weaker or inactive substrate for the RNA Ligase in the next cycle.

Terminal deoxynucleotidyl Transferase (TdT) is incapable of coupling its corresponding 3'-phosphate nucleotide substrate analog, dNTP-3'-phosphate. A suggestion has been made in the literature for producing a mutant form of TdT capable .of coupling dNTP-3'-phosphate. (Chang et al, *CRC Critical Reviews in Biochemistry,* 21(1): 27–52). Such a mutant form would be a useful chain extending enzyme for the method of the invention.

NUCLEOTIDE SUBSTRATES

The blocked nucleotide substrate employed in the method of the invention is selected for compatibility with the chain extending enzyme, but generally comprises an activated nucleotide and a blocking group. The blocking group is bonded to the nucleotide so as to block reaction of the 3'-hydroxyl group of the nucleotide. Such a nucleotide substrate is referred to generally herein as a "3'-blocked nucleotide."

As used herein, the term "3'-phosphate-blocked nucleotide" refers to nucleotides in which the hydroxyl group at the 3'-position is blocked by the presence of a phosphate containing moiety. Examples of 3'-phosphate-blocked

nucleotides in accordance with the invention are nucleotidyl-3'-phosphate monoester/nucleotidyl-2',3'-cyclic phosphate, nucleotidyl-2'-phosphate monoester and nucleotidyl-2' or 3'-alkylphosphate diester, and nucleotidyl-2' or 3'-pyrophosphate. Thiophosphate or other analogs of such compounds can also be used, provided that the substitution does not prevent dephosphorylation by the phosphatase.

When RNA ligase is employed as the chain extending enzyme, the choice of substrate influences the course of the reaction, as can be seen from a consideration of the following reaction mechanism:

(1) E+ATP ⟷ E-AMP+pyrophosphate

(2) E-AMP+3',5'-(d)NDP ⟷ E[App(d)Np]

(3) E[App(d)Np]+primer-3'-OH ⟷ (primer-p (d)N)-3'-phosphate+AMP+E wherein App is an adenosine diphosphate moiety and Np is a 3'-phosphate blocked nucleoside moiety, preferably a 3'-phosphate monoester. The use of precursor nucleotides, ATP+3',5'-(d)NDP, results in a short lag period in the coupling reaction in which the concentration of App(d)Np must build up to sufficient levels in solution before step **3** can occur. The use of pre-activated nucleotide substrate, App(d)Np, avoids a lag period, allowing step **3** to occur instantly. Therefore, faster and more reliable RNA ligase coupling can be achieved using pre-activated nucleotide substrates.

The scientific literature documents that the adenylylated enzyme is unable to catalyze step **3** of the reaction. The addition of a small amount of 3',5'-(d)NDP, when using pre-activated nucleotide substrate, App(d)Np, is believed by the inventor to prevent RNA ligase from being irreversibly inactivated by the reverse reaction of step **2**. Consequently, it is believed that the coupling reaction proceeds with greater efficiency. The addition of a small amount of pyrophosphate may perform the same function.

Pre-activated blocked nucleotides for use as substrates in the method of the invention can be conveniently synthesized in accordance with Example 1.

Other substrates which are coupled to the primer by the chain extending enzyme and which can be converted to an inert or slowly reacting product may also be employed.

DEBLOCKING ENZYMES

When the 3'-blocking group employed on the substrate is a phosphate group, the enzyme employed to remove the blocking group is a phosphatase. The principal function of the phosphatase is the irreversible removal of the 3'-phosphate blocking group from the extended primer (allowing subsequent RNA ligase coupling) and optionally, removal from the nucleotide substrate (preventing subsequent RNA ligase coupling). Careful selection of the phosphatase and the reaction conditions allows either: (1) dephosphorylation of both the extended primer and unreacted nucleotide substrate when substrate is not to be reused; or (2) dephosphorylation of only the extended primer when substrate is to be reused in the next cycle. Non-specific phosphatases such as Alkaline Phosphatase and Acid Phosphatase are useful when substrate reuse is not desired, as depicted in FIG. **2**A; specific 3'-Phosphatases such as T4 3'-Phosphatase and Rye Grass 3'-Phosphatase are useful when substrate reuse is desired.

Alkaline Phosphatase will hydrolyze any monoester phosphate. Its high activity, especially at elevated temperatures, its substantial inability to degrade oligonucleotides, and its ability to be denatured irreversibly at 95° C. make it a useful deblocking enzyme in the invention. Alkaline phosphatase is readily available commercially from intestine and from bacteria. The inherent inorganic pyrophosphatase activity of alkaline phosphatase, not present in T4 3'-phosphatase, prevents a pyrophosphate build-up which may inhibit RNA ligase.

Acid Phosphatase has been isolated from wheat, potato, milk, prostate and semen, and catalyzes the same reactions as Alkaline Phosphatase. Acid Phosphatase can substitute for Alkaline Phosphatase if the pH of the reaction solution is acidic. Alkaline phosphatase is the preferred deblocking enzyme, however, when substrate is not to be reused in the next cycle.

The 3'-Phosphatases can be used either to dephosphorylate the primer selectively or to dephosphorylate both the primer and the nucleotide substrate depending on the reaction conditions selected. Low concentrations are used for selective dephosphorylation; high concentrations are used to dephosphorylate both.

The technical challenge of selective dephosphorylation is that it entails removal of the blocking group from the primer-blocked nucleotide product without removal of the blocking group from the unreacted blocked nucleotide substrate. In the method of the invention using RNA Ligase as the chain extending enzyme and AppNp as nucleotide substrate, the technical difficulty is selectively removing the 3'-phosphate blocking group of the extended primer, primer-pN-3'-phosphate, without removing the 3'-phosphate of the nucleotide substrate AppN-3'-phosphate. This difficulty is exacerbated by the fact that primer-pN-3'-phosphate and AppN-3'-phosphate are structurally identical with respect to the 3'-phosphate group in that they both share the same pN-3'-phosphate unit; the structural difference exists in a region distant from the 3'-phosphate, the component connected to the 5'-phosphate. This high degree of structural similarity would seemingly make discriminating between the substrates unachievable. Furthermore, the degree of discrimination (selectivity) must be sufficiently high to make a nucleotide substrate reuse technique useful. In the present invention, this challenge is solved as a result of the discovery that the enzyme 3'-Phosphatase is capable-of achieving the selective dephosphorylation and that it does so in a manner which makes the invention useful.

3'-Phosphatase dephosphorylates only 2'- or 3'-phosphate esters. Two 3'-Phosphatases are commercially available: bacteriophage T4 and rye grass; both are useful in the method of the invention. The T4 enzyme is a bifunctional enzyme containing Polynucleotide Kinase and 3'-Phosphatase activities, catalyzed from two independent active sites. The T4 enzyme is commonly sold as "Polynucleotide Kinase". Since it is the 3'-phosphatase activity which is of main relevance in this invention, this enzyme herein will be referred to as T4 3'-Phosphatase. 3'-Phosphatase derived from rye grass is sold commercially as "3'-Nucleotidase" (Sigma Chemical, E. C. 3.1.3.6). This enzyme will also herein be referred to in this specification as 3'-Phosphatase. The method of the invention embodies any 3'-Phosphatase with the aforementioned substrate selectivity.

Genetic mutants of T4 3'-Phosphatase which lack associated kinase activity would also be useful in the invention. This task has already been described in the literature. A genetic mutant called pseT**47** and a proteolytic fragment of the enzyme have the 3'-Phosphatase activity, but no kinase activity. Soltis et al., *J. Biol. Chem.* 257:11340–11345 (1982). Removal of the associated kinase activity may be desirable in preventing oligonucleotide circularization or polymerization. Other useful 3'-Phosphatases may be constructed by making genetic mutations which remove undesirable associated enzyme activities.

Given that 3'-Phosphatase is probably widespread in nature, it is anticipated that other 3'-Phosphatases derived from other sources will display similar or perhaps superior selective dephosphorylation and will also be useful in the invention. Thus far, experiments performed by the inventor have been unable to demonstrate that reuse of substrates can be applied to deoxyribose substrates AppdNp, since it appears that 3'-Phosphatase lacks the ability to selectively dephosphorylate primer-pdNp without substantially dephosphorylating AppdNp. A corresponding 2'-deoxy-3'-phosphatase with the aforementioned selectivity would be useful for AppdNp substrate reuse.

Special consideration is necessary for the method of FIGS. **3**A and B to avoid significant co-incubation of 3'-phosphatase activity and RNA ligase activity in the presence of primer+AppNp, which may result in uncontrolled substrate addition. For example, RNA ligase may be heat inactivated after use, or using a thermostable enzyme, the RNA ligase activity can be temporarily turned off by lowering the temperature during the 3'-phosphatase incubation.

T4 3'-phosphatase has potential disadvantages with respect to its use in the synthesis of non-repeat regions of an oligonucleotide, as follows: (1) The 3'-phosphatase activity on unreacted nucleotide substrate is substantially slower than Alkaline Phosphatase; (2) AMP which is generated by the RNA ligase coupling reaction is not hydrolyzed by 3'-phosphatase and its accumulation after many coupling cycles may inhibit RNA ligase; and (3) associated kinase activity may result in cyclization or polymerization of the oligonucleotide if ATP is employed in the RNA ligase coupling reaction. Thus, while T4 3'-phosphatase is useful for all aspects of the method of the invention, the preferred Phosphatase for synthesis of non-repeat regions is Alkaline Phosphatase.

Other blocking groups which might be used in the method of the invention include blocking groups which are removed by light, in which case the addition of an enzyme to accomplish the unblocking would be unnecessary. See Ohtsuka et al, *Nucleic Acids Res,* 6(2):443–54 (1979). Other blocking groups include any chemical group covalently attached to the 2'- or 3'-hydroxyl of App(d)N-3'-OH, which can be removed without disrupting the remainder of the oligonucleotide. This may include esters, sulfate esters, glucose acetals, a heat labile group, or an acid or base labile group, which can be removed by incubation with esterases or proteases, sulfatases, glucosidases, heat, or acid or base, respectively.

ADDITIONAL METHOD STEPS

To synthesize long oligonucleotides, it is desirable to overcome two potential problems: the extension of the chain with unreacted nucleotide of the wrong type, and the subsequent extension of failed reaction products (unextended primer) from a previous cycle. These problems can be overcome by the addition of one or more additional enzymes to the basic scheme shown in FIG. **2**A or **3**A.

When synthesizing long oligonucleotides, such as about 25 bases or more, the unblocked nucleotide App(d)N concentration may build up to an extent that it couples to the primer at an unacceptable level, despite the fact that it is far less reactive than App(d)Np substrate. To minimize the incorporation of such residual nucleotides from previous reaction cycles, an additional enzyme can be added during, after or prior to the unblocking step which is effective to further degrade unreacted nucleotide substrate or nucleotide fragments into products that are no longer suitable substrates for RNA ligase.

A suitable enzyme for this purpose is a Dinucleotide Pyrophosphate Degrading Enzyme. Five distinct enzymes are capable of degrading App(d)N or App(d)Np, as described in the scientific literature:

(1) Nucleotide Pyrophosphatase (E. C. 3.6.1.9)

(2) Acid Pyrophosphatase (Tobacco, Sigma Chemical Co.)

(3) Diphosphopyridine Nucleosidase (E. C. 3.2.2.5)+ADP-Ribose Pyrophosphatase (E. C. 3.6.1.13)

(4) Dinucleotide Pyrophosphate Deaminase (Kaplan et al, *J. Biol. Chem.,* 194: 579–91 (1952))

(5) Dinucleotide Pyrophosphate Pyrophosphorylase (A. Kornberg, *J. Biol. Chem.,* 182: 779–93 (1950))

These enzymes are suitable for this invention because the degradation products are not substrates for RNA ligase. Among the Dinucleotide Pyrophosphate Degrading Enzymes, the preferred enzyme is Nucleotide Pyrophosphatase. This enzyme offers the following advantages: the reaction is irreversible, the enzyme degrades both App(d)N and App(d)Np; and nucleotide substrate is hydrolyzed to nucleosides+PO_4 when used with Alkaline Phosphatase. This is advantageous since nucleosides and phosphate are substantially non-inhibitory to all the enzymatic reactions of the method. Precipitation of nucleosides as a result of accumulation and poor solubility is probably beneficial by making the nucleosides inert to all reactions of the oligonucleotide synthesis, and by facilitating separation of the nucleosides from the final oligonucleotide product by centrifugation. The use of App(d)N_1p(d)N_2p as a nucleotide substrate for RNA ligase requires the use of a Dinucleotide Pyrophosphate Degrading enzyme and Alkaline Phosphatase to achieve inactivation for use in the method.

Nucleotide Pyrophosphatase has been isolated from a great number of sources: human fibroblasts, plasmacytomas, human placenta, seminal fluid, *Haemophilus influenzae,* yeast, mung bean, rat liver, and potato tubers. The source with the best characterized enzymatic properties is potato tubers Bartkiewicz et al, *Eur. J. Blochem.,* 143:419–26 (1984). Bartkiewicz et al have shown that purified enzyme is capable of hydrolyzing dinucleotide pyrophosphates specifically, without hydrolyzing DNA or RNA. Nucleotide Pyrophosphatase isolated from snake venom is commercially available (Sigma Chemical Co.) and is the same enzyme as Phosphodiesterase I. (PDE-I) Accordingly, PDE-I can also be used to convert unreacted nucleotides into a form which does not serve as a substrate for RNA ligase, However, the exonuclease activity of the PDE-I warrants careful consideration since this activity may destroy the oligonucleotide product and prior art does not teach the use of exonucleases in a synthetic method.

The second potential difficulty with the method of the invention arises from a build up of failure sequences due to incomplete RNA ligase coupling. The RNA ligase coupling reaction can be substantially optimized kinetically in accordance with the invention. Dithiothreitol and TRITON X-100 (octylphenoxy polyethoxy ethanol) greatly stimulate RNA ligase activity. Nevertheless, even under optimized conditions, the coupling reaction is not 100% efficient, resulting in primer chains which have not been coupled to the blocked nucleotide. If not removed, these unreacted primer chains will still be able to couple with nucleotide in the next coupling cycle. This will result in the accumulation of (n–1) failure sequences in the final product mix. Two independent solutions have been devised-by the inventor to solve this problem: Exonuclease treatment and Enzymatic Capping.

An exonuclease can be added after RNA ligase coupling to hydrolyze uncoupled primer chains to (d)NMP's. The Exonuclease can be utilized before, after, or concurrently with the dinucleotide pyrophosphate degrading enzyme. The Exonuclease used for this purpose should have the following properties:

(1) hydrolyzes oligonucleotides in the 3' to 5' direction; and

(2) hydrolyzes specifically oligonucleotides with a free terminal 3'-hydroxyl group and is substantially unable to hydrolyze oligonucleotides which are blocked at the 3'-end.

Primer chains which fail to couple during incubation with RNA ligase differ from primer chains which do couple. Uncoupled primers have a 3'-hydroxyl terminus; coupled primers have a blocked 3'-phosphate. Therefore, as a result of the selectivity of the Exonuclease, only uncoupled primer chains are degraded to (d)NMP's. Exonuclease incubation should be performed prior to incubation with Phosphatase, and exonuclease activity should not be present during phosphatase incubation. Otherwise, oligonucleotide product will be hydrolyzed.

Three enzymes satisfy these criteria and are suitable as Exonuclease in this invention: Exonuclease I (*E. coli*), Phosphodiesterase I (snake venom), and Polynucleotide Phosphorylase. Phosphodiesterase I hydrolyzes both oligoribonucleotides and oligodeoxyribonucleotides; Exonuclease I is substantially specific for oligodeoxyribonucleotides (although it has been used successfully on mixed deoxyribose/ribose oligonucleotides); Polynucleotide Phosphorylase is substantially specific for oligoribonucleotides. TRITON X-100 and dithiothreitol have been observed experimentally by the inventor to stimulate the activity of Exonuclease I and PDE-I.

PDE-I offers two advantages: (1) PDE-I hydrolyzes both oligoribonucleotides and oligodeoxyribonucleotides, making it useful for the synthesis of both, and (2) PDE-I has nucleotide pyrophosphatase activity. Although PDE-I requires careful control of enzymatic reaction conditions to avoid degrading primer chains blocked by a 3'-phosphate, conditions can be achieved to hydrolyze all 3'-hydroxyl primer chains and all unreacted blocked nucleotide substantially without hydrolyzing 3'-phosphate primer chains. Given that it is advantageous to use a Dinucleotide Pyrophosphate Degrading activity and an Exonuclease activity simultaneously, snake venom PDE-I provides two functions for the price of one enzyme.

The combination of these two modifications of the Basic method results in the Preferred method for the synthesis of oligonucleotides, outlined in FIGS. **2B** and **3B**. The power of this method is exemplified in Example 5. ApApCpdApdA is synthesized by two coupling cycles with the activated nucleotide AppdAp and ApApCp initial primer. Thin layer chromatography demonstrated that the reaction mixture at the end of the synthesis contained only the oligonucleotide ApApCpdApdA and the nucleosides adenosine and deoxyadenosine. The mixture was devoid of traces of n–1 and n–2 failure sequences. Due to the enormous size difference between the n-mer oligonucleotide product and the nucleosides, the oligonucleotide product can be easily purified. Furthermore, an application may not require removal of the nucleosides.

As mentioned earlier, another technique can be used to remove uncoupled primer chains, denoted herein as "Enzymatic Capping." After the RNA ligase coupling reaction, unreacted primer chains can be capped with a chain terminating nucleotide catalyzed by a transferase enzyme. The capped chains are no longer substrates for coupling with RNA ligase in subsequent coupling cycles. Primer chain termination can be achieved with Terminal deoxynucleotidyl Transferase+dideoxynucleoside triphosphate or with RNA ligase+AppddN (the dideoxy analog of AppdN). Chain terminated failure sequences can be subsequently hydrolyzed to nucleotides using an exonuclease as described above. One potential disadvantage of the enzymatic capping technique is the coupling efficiency of the chain terminating step. If the coupling efficiency is low, then (n–1) failure sequences will be present in the final solution mixture. Thus, the favored method for removing uncoupled primer chains is the Exonuclease method discussed earlier.

REMOVAL OF AMP

AMP generated during the coupling reaction may inhibit the forward coupling reaction or participate in the reverse coupling reaction. In accordance with the invention, this can be avoided by the addition of an enzyme or enzyme combination which degrades AMP to a less inhibitory form. For the purpose of this invention, an AMP Inactivating Enzyme or Enzyme Combination, is defined as an enzyme or enzyme combination which converts Adenosine 5'-Monophosphate (AMP) to a less reactive form, i.e., to a form which is less inhibitory to the forward coupling reaction catalyzed by RNA Ligase, or which is less able to participate in the reverse coupling reaction catalyzed by RNA Ligase, or which assists in driving (thermodynamically or kinetically) the forward coupling reaction catalyzed by RNA Ligase. An AMP Inactivating Enzyme or Enzyme Combination is useful in making the RNA Ligase coupling reaction faster, more efficient, or more reliable, by converting AMP, generated by the forward coupling reaction, to a form with diminished undesirable properties.

Several AMP Inactivating Enzymes have been devised by the inventor. These enzymes are preferably used concurrently with RNA Ligase incubation since they do not substantially degrade primer, extended primer product, or App-(d)Np substrate. These enzymes can be present or can be used at any or all steps of a cycle since their activity is not deleterious to the One Pot method. Such enzymes include:

(1) 5'-Nucleotidase (E. C. 3.1.3.5):
$AMP+H_2O \rightarrow Adenosine+phosphate$

(2) AMP Nucleosidase (E. C. 3.2.2.4):
$AMP+H_2O \rightarrow Adenine+ribose\text{-}5\text{-}phosphate$

(3) AMP Deaminase (E. C. 3.5.4.6):
$AMP+H_2O \rightarrow Inosine\text{-}5'\text{-}phosphate+NH_3$

For clarity, FIG. **8** shows the structure of AMP and the location of the covalent bond broken by the hydrolytic activity of each enzyme. Experiments by the inventor strongly suggest that the hydrolytic products of these enzymes are less inhibitory to RNA Ligase than AMP. Furthermore, it is strongly suspected that these hydrolytic products are unable to participate in the reverse RNA Ligase coupling reaction. Example 19 demonstrates the use of these enzymes.

These three enzymes using AMP substrate may be combined in a rational manner with other enzymes, which further convert their products to even less reactive products, to create an AMP Inactivating Enzyme Combination. Such enzymes include:

(1) Adenosine Nucleosidase (E. C. 3.2.2.7):
$Adenosine+H_2O \rightarrow Adenine+ribose$

(2) Adenosine Deaminase (E. C. 3.5.4.4):
$Adenosine+H_2O \rightarrow Inosine+NH_3$

(3) Nucleoside Phosphorylase (E. C. 2.4.2.1):

Adenosine+PO_4→ribose-1-phosphate+Adenine

Example **19** demonstrates the enzyme combination 5'-Nucleotidase+Adenosine Deaminase. Other potentially useful combinations, such as 5'-Nucleotidase+Adenosine Nucleosidase, can be constructed by identifying the side product which one wishes to convert to a less reactive form and consulting *Enzyme Nomenclature* (Academic Press, 1992) or the scientific literature to locate an enzyme which effects the conversion. For example, to remove adenine., consultation with *Enzyme Nomenclature* discloses the enzyme Adenine Deaminase (E. C. 3.5.4.2) which converts adenine to hypoxanthine, which may be suitable for inclusion in an enzyme combination. Similarly, uridine can be converted to uracil by adding Uridine Nucleosidase (E. C. 3.2.2.3).

AMP Nucleosidase and AMP Deaminase are reported in the literature as allosterically activated by ATP and allosterically inactivated by phosphate. Experiments indicate that these enzymes have adequate activity in the absence of ATP and under the conditions employed for oligonucleotide synthesis demonstrated in Example **19**. A thermostable version is probably obtainable from a thermophilic organism, e.g. *Thermus aquaticus, Pyrococcus,* etc and would be useful in the method since replenishment would be unnecessary.

The concept of an AMP Inactivating Enzyme or Enzyme Combination as a useful technique in the method of the invention is not limited to the enzymes disclosed in this specification, but shall include any enzyme which can be implemented for the previously stated purpose. Such enzymes may already be described in the literature, may be discovered in the future, or may be a man-made genetic modification of a known AMP Inactivating Enzyme. For example, a mutant of either AMP Nucleosidase or AMP Deaminase with constitutively high activity would be useful. Many examples exist in the literature in which mutations affect allosteric enzyme properties.

SUPPLEMENTAL TECHNIQUES

While the foregoing describes the basic aspects of the claimed invention, it will be appreciated that numerous modifications are possible without departing from the basic invention.

In practicing the method of the invention, enzyme inactivation where needed can be readily accomplished using heat or by proteolysis with a protease, e.g., proteinase K. Protease can be subsequently inactivated by heat or by chemical inhibitor such as phenylmethylsulfonyl chloride.

Proteolysis with proteinase K can also be used to hydrolyze the denatured protein debris, which accumulates as a result of heat inactivation of enzymes, to small soluble peptides. Although the debris is inert, its accumulation after many cycles may pose a viscosity problem for mixing or pipetting operations. The proteolytic digestion may be enhanced by the addition of TRITON X-100. Physical methods for removing the debris such as filtration, ultrafiltration, centrifugation, and extraction with organic solvents such as phenol and chloroform can also be utilized, but are not readily automated and are more appropriate as an option at the end of the synthesis.

The method of the invention is particularly well adapted to the synthesis of oligoribonucleotides. It can also be used to synthesize oligodeoxyribonucleotides, although coupling times will be longer and coupling efficiencies will be lower. For most applications an oligoribonucleotide can substitute for an oligodeoxyribonucleotide with equal effectiveness. Oligoribonucleotides can be used as hybridization probes, as primers for dideoxy DNA sequencing (RNase can remove the primer prior to electrophoresis); as primers for the polymerase chain reaction using a thermostable reverse transcriptase; and as probes for the ligase chain reaction.

For applications which have an absolute requirement for oligodeoxyribonucleotides, an oligoribonucleotide may be converted to its complementary oligodeoxyribonucleotide. The oligoribonucleotide can be synthesized with a hairpin at the 3'-end, allowing priming for reverse transcriptase, and subsequent RNase H digestion.

Large scale manufacture of enzymes employed in the present invention having suitable purity may be accomplished by established methods for expression of recombinant protein in an overproducing organism. One such technique is to manufacture the enzymes as fusion proteins with an affinity protein, allowing purification in one step by affinity chromatography. As the activity of many enzymes is not affected by the presence of the affinity protein, proteolytic removal of the affinity protein is probably not necessary.

Alternative embodiments of the present invention may be implemented to reduce the cost of enzymes. For example, instead of inactivating the enzymes by heat or proteolysis, enzymes may be recovered from the oligonucleotide solution by passing the solution through an enzyme-binding solid support, such as an affinity chromatography column, and then optionally reused in later cycles of the invention. Alternatively, enzymes may be covalently attached to a solid support matrix and placed in columns. The method of the invention is then performed by pumping solutions through the appropriate columns.

The hydrolysis of phosphate anhydrides by Alkaline Phosphatase and Nucleotide Pyrophosphatase, and the hydrolysis of phosphodiesters by Exonuclease releases an equivalent of acid. Preventing an unacceptable drop in pH, especially for long oligonucleotides, may entail the occasional addition of base or the use of a higher buffer concentration.

Phosphate concentrations exceeding about 20 mM at pH 8.0 and 10 mM $MgCl_2$ may eventually precipitate the magnesium. This is deleterious since magnesium is a required cofactor for many of the enzymes in the One Pot method. This problem can be solved by conducting the synthesis at pH 7.0. Experiments confirm that no precipitation of $MgPO_4$ is observed in a solution of 10 mM $MgCl_2$ and 250 mM PO_4 at pH 7.0. Alternatively, phosphate can be removed by precipitation out of solution by adding an excess of Mg^{++}, Ca^{++}, Al^{+++} or other cationic species which forms an insoluble phosphate salt. Hydrolysis of pyrophosphate by Inorganic Pyrophosphatase prevents precipitation of magnesium pyrophosphate, which is highly insoluble in aqueous solutions.

Growth in the reaction mixture of microorganisms may result in the secretion of nucleases which could degrade the nucleotides and oligonucleotides. This problem is minimized by the frequent heat inactivation steps which sterilize the reaction solution and the use of the detergent TRITON X-100 which may hinder most microbes. Alternatively, microbial growth inhibitors, such as glycerol, EDTA, sodium azide, merthiolate, or antibiotics may be added to the reaction solution. A useful growth inhibitor for the method of the invention should not significantly inhibit the enzymatic reactions in the synthesis of the oligonucleotide. No significant inhibition of RNA ligase was observed by the

inventor in the presence of 0.1% sodium azide and 0.1% merthiolate.

Inadvertent nuclease contamination of the synthesis reaction can be countered by adding a nuclease inhibitor or adding protease intermittently. Numerous RNase inhibitors are described in the literature,. including RNase Inhibitor Protein (human placenta) and Vanadyl Ribonucleoside Complexes (Sigma Chemical Co). No significant inhibition of RNA ligase was observed by the inventor in the presence of 0.1 mM vanadyl ribonucleoside complexes.

Evaporative loss can be minimized by reducing the temperature or duration of the heat inactivation steps or by overlaying the aqueous phase with light mineral oil. For example, snake venom PDE-I can be inactivated by heating at 50° C. for 5 minutes; commercially available heat labile alkaline phosphatase from Arctic fish can be inactivated at 65° C.

Consumption of dithiothreitol, or other reducing agents, that stimulate the activity of the enzymes used in the One Pot method by oxidation may be solved by either intermittent replenishment or by conducting the synthesis in an oxygen-free environment.

The formation of secondary structure in an oligonucleotide may block enzymatic access to the 3'-end of the oligonucleotide. Several measures may be taken. The oligonucleotide can be synthesized as several smaller pieces which do not self anneal and then ligated together with RNA ligase. Alternatively, the base portion of a nucleotide can be modified with protecting groups such as acetyl groups which prevents base pairing. The protecting groups are removed at the end of the synthesis. A third alternative is the addition of denaturants to the reaction mixture which disrupt oligonucleotide base pairing without substantially inhibiting the enzymatic reactions. Suitable denaturants include dimethyl sulfoxide, formamide, methylmercuric hydroxide and glyoxal. No significant inhibition of RNA ligase was observed by the inventor in the presence of 20% dimethyl sulfoxide.

APPARATUS

The minimal configuration for an apparatus which is useful for synthesizing oligonucleotides by the method of the invention is: (1) at least one vessel containing reaction solution for performing the synthesis of an oligonucleotide, (2) means for controlling the temperature of the reaction solution(s), (3) means for separately supplying at least four different blocked nucleotide feed stocks to the solution(s), (4) means for supplying at least one enzyme feed stock to the solution(s), and (5) means for controlling the sequential addition of blocked nucleotide feed stocks and enzyme feed stock(s) to the solution(s). Two separate embodiments of the minimal configuration are described.

FIG. **9** shows an apparatus which can be used in the practice of the invention for synthesizing many oligonucleotides simultaneously. The apparatus has a plurality of reaction vessels in the form of wells **2** drilled in a metal block **1**. At least four different blocked nucleotide feed stocks and at least one enzyme feed stock are provided from reagent bottles **4** using one or several liquid handling robots **3**. The temperature of the block can be increased by turning on a heating element (not shown) beneath the block and can be lowered by opening a valve **6** which allows water **5** to flow through a cavity (not shown) underneath the block and then exit **7**. A computer (not shown) controls the sequential addition of blocked nucleotides and enzyme(s) to the vessels and controls the temperature of the block.

This apparatus can be further improved by providing a separate means for mixing the synthesis reaction solutions without the need for the robotic liquid dispensing system to mix reaction solutions. This can be accomplished by placing a magnetic stir bar or many small magnetic or paramagnetic particles in each of the wells in active use, and agitating the stir bars with a moving magnetic field. Wells may be coated with an inert material to avoid heavy metal contamination.

FIG. **10** shows an apparatus which can be used in the practice of the invention for synthesizing a single oligonucleotide in bulk quantity. It consists of a single large vessel **53** for the synthesis reaction which is mixed by a stirring device. The stirring device may be a motor **51** connected to a rotating impeller **52**, or alternatively a large stir bar (not shown) rotated by a magnetic stirrer (not shown). The temperature of the reaction solution is increased with a heating device **54** or a heating element (not shown) located inside cavity **60**, and lowered by opening a valve **59** which allows cool water **58** to flow into a cavity **60** beneath the vessel and then exit **61** the cavity. The four blocked nucleotide feed stocks **63** are added to the vessel either by four separate pumps (not shown) or by a single pump with a valve controlling connection of the feed stocks to the pump (not shown). At least one enzyme feed stock **64** can be added in the same manner. A computer (not shown) controls the sequential addition of blocked nucleotides and enzyme(s) to the vessel and controls the temperature of the solution.

Additional components could enhance the performance of the bulk scale synthesizer. Ancillary feed stocks **65** for additional blocked nucleotides, enzymes, or other reagents can be added. The temperature of the reaction solution is monitored by a temperature probe **55**. A pH probe **56** monitors the reaction solution pH and acid or base feed stocks **62** can be added as necessary to maintain pH as desired. An inert gas such as nitrogen is slowly added via tube **57** to the reaction solution to remove oxygen (which can be monitored by an oxygen electrode). A computer (not shown) can control the apparatus, receiving inputs of solution temperature, pH, and sending outputs to control the addition of feed stocks (blocked nucleotide feed stocks, enzyme feed stock(s), acid, base, and ancillary reagents), heating device, cooling valve **59**, nitrogen purge rate, and motor rotation speed. Nucleoside and phosphate by-products may be reduced by adding a dialysis or ultrafiltration system (not shown).

Reagents

Several reagents useful in the practice of the invention have not been previously described, and these reagents are an aspect of the present invention. In particular, the activated deoxyribonucleotides AppdAp, AppdGp and AppdCp; and dinucleotides of the general formula $App(d)N_1p(d)N_2p$, wherein N_1 and N_2 are any nucleosides.

The activated deoxyribonucleotides can be synthesized by phosphorylation of the 5'-hydroxyl of the corresponding 3'-dNMP using phosphatase free polynucleotide kinase and ATP, to yield 3',5'-dNDP. This is then activated in accordance with Example 1.

The dinucleotides can be synthesized in several steps. First, $(d)N_1p(d)N_2p(d)N_3$ is synthesized chemically, for example using the phosphoramidite method. This product is then phosphorylated using ATP and Polynucleotide Kinase to yield $p(d)N_1p(d)N_2p(d)N_3$. The enzyme is then inactivated. The phosphorylated material is then partially

digested, e.g. using RNase, DNase or a nuclease to yield $p(d)N_1p(d)N_2p$. The enzyme activity is then removed using protease followed by heat, after which the material is activated as in Example 1. Activation of such dinucleotides substrates is greatly accelerated by the presence of a primer.

While the method of invention can be described in terms of a cycle of steps which result in synthesis of oligonucleotides, certain aspects of the invention are independently viewed as part of applicant's inventive concept. For example, the application of an exonuclease to degrade any oligonucleotide primer which was not extended is a useful improvement in the context of any method for synthesizing an oligonucleotide, wherein an oligonucleotide primer is extended coupling a blocked nucleotide to the 3'-end of the primer, wherein primer-blocked nucleotide product is resistant to exonuclease attack. Similarly, the application of a transferase enzyme and a chain terminating nucleotide, whereby any oligonucleotide primer which was not extended is end-capped to render it unreactive to further extension in any method for synthesizing an oligonucleotide is a useful improvement in the context of any method for synthesizing an oligonucleotide wherein an oligonucleotide primer in a reaction mixture is extended by coupling a blocked nucleotide to the 3'-end of the primer, such that primer-blocked nucleotide product is formed that is unreactive with the transferase enzyme.

The method will now be further described by way of the following, non-limiting examples.

EXAMPLE 1

Enzymatic Synthesis of AppAp and AppdAp

The synthesis of activated nucleotides, App(d)Np and App(d)Np(d)Np can be performed enzymatically using RNA ligase Inorganic Pyrophosphatase. This example demonstrates the synthesis of AppAp and AppdAp; other activated nucleotides can be synthesized in the same manner.

The following solution in a total volume of 300 ul was placed in an ependorf tube: 50 mM Tris-Cl, pH 8.0, 10 mM $MgCl_2$, 10 mM Dithiothreitol (DTT), 0.1% TRITON X-100, 11 mM 3',5'-ADP, 10 mM ATP, 0.1 units Inorganic Pyrophosphatase (yeast, Sigma Chemical Co.), 80 units RNA ligase (phage T4, New England Biolabs). For the synthesis of AppdAp, 3',5'-dADP was used in place of 3',5'-ADP. This solution was incubated at 37° C. for 40 hours. RNA ligase was heat inactivated at 95° C. for 5 minutes. Residual ATP was removed by adding 2 units Hexokinase (yeast, Sigma Chemical Co.)+15 ul 200 mM glucose and incubating at 37° C. for 1 hour. Hexokinase was heat inactivated at 95° C. for 5 minutes. The solution was cooled to room temperature and pelleted at 12,000 g for 1 minute to remove the insoluble protein debris. This final product was analyzed by thin layer chromatography on silica using isobutyric acid:concentrated ammonium hydroxide:water at 66:1:33 containing 0.04% EDTA (hereinafter "butyric-TLC"). No ATP was detected; the major product was App(d)Ap with a small amount of 3',5'-ADP present. AppAp and AppdAp prepared in this manner were used in all the following examples.

EXAMPLE 2

One Pot Synthesis of ApApCpApA

The following solution was placed in a total volume of 40 ul in an ependorf tube: 50 mM Tris-Cl, pH 8.0, 10 mM $MgCl_2$, 10 mM DTT, 0.1% TRITON X-100, 1 mM ApApC primer, 5 mM AppAp. The following procedures were performed:

cycle 1

(a) Add 2 ul (40 units) RNA ligase (phage T4, New England Biolabs). Incubate at 37° C. for 15 minutes. Heat at 95° C. for 5 minutes, cool to room temperature.

(b) Add 1 ul (3 units) Alkaline Phosphatase (calf intestine, US Biochemicals). Incubate at 37° C. for 30 minutes. Heat at 95° C. for 5 minutes, cool to room temperature.

cycle 2 - starting volume is 20 ul

(a) Add 10 ul 10 mM AppAp+1 ul RNA ligase. Incubate at 37° C. for 30 minutes. Heat at 95° C. for 5 minutes, cool to room temperature.

(b) same as cycle 1

Insoluble coagulated protein-debris was removed by pelleting at 12,000 g for 1 min. The reaction mixture supernatant was analyzed by thin layer chromatography using the SureCheck™ Oligonucleotide Kit (US Biochemicals)(hereinafter "USB TLC"). The only oligonucleotide product visible on the TLC plate was the desired oligonucleotide product ApApCpApA; i.e., no n–2, n–1, n+1, n+2, etc. products were formed. This experiment demonstrates that AppA does not participate in the RNA ligase coupling reaction, due to its slow coupling rate relative to AppAp. This experiment also demonstrates that coupling times with efficiencies approaching 100% can be achieved in 15 minutes under these experimental conditions. This is attributable to the nucleotide 3',5'-ADP present in the AppAp preparation, which prevents covalent inactivation of RNA ligase. The final yield of oligonucleotide product approached 100%.

EXAMPLE 3

Synthesis of (ApApC)-pApA

The following solution was placed in a total volume of 30 ul in an ependorf tube: 50 mM Tris-Cl, pH 8.0, 10 mM $MgCl_2$, 10 mM DTT, 0.1% TRITON X-100, 1 mM ApApC primer, 5 mM AppAp. The following procedures were performed:

cycle 1

(a) Add 1 ul (20 units) RNA ligase (phage T4, New England Biolabs). Incubate at 37° C. for 1 hour. Heat at 95° C. for 5 minutes, cool to room temperature.

(b) Add 1 ul (0.03 units) Nucleotide Pyrophosphatase (snake venom, Sigma Chemical Co. P7383). Incubate at 37° C. for 30 minutes. Heat at 95° C. for 5 minutes, cool to room temperature.

(c) Add 1 ul (3 units) Alkaline Phosphatase (calf intestine, US Biochemicals). Incubate at 37° C. for 30 minutes. Heat at 95° C. for 5 minutes, cool to room temperature.

cycle 2

(a) Add 10 ul 10 mM AppAp+1 ul RNA ligase. Incubate at 37° C. for 5 hours. Heat at 95° C. for 5 minutes, cool to room temperature.

(b) same as cycle 1

(c) same as cycle 1

Insoluble coagulated protein debris was removed by pelleting at 12,000 g for 5 min. USB TLC revealed pure ApApCpApA product with no visible n–1 or initial primer present. The yield of final product was about 90% of the initial primer.

EXAMPLE 4

Synthesis of (ADApC)-pADA

The following solution was placed in a total volume of 30 ul in an ependorf tube: 50 mM Tris-Cl, pH 8.0, 10 mM

$MgCl_2$, 10 mM DTT, 0.1% TRITON X-100, 1 mM ApApC primer, 5 mM AppAp. The following procedures were performed:

cycle 1

Performed identically to cycle 1 of example 3.

cycle 2

(a) Add 1.5 ul 100 mM ATP+3 ul 50 mM 3'5'-ADP+0.1 units Inorganic Pyrophosphatase+1 ul RNA ligase. Incubate at 37° C. for 5 hours. Heat at 95° C. for 5 minutes, cool to room temperature.

(b) same as cycle 1 of example 3

(c) same as cycle 1 of example 3

Insoluble coagulated protein debris was removed by pelleting at 12,000 g for 5 min. USB TLC revealed nearly pure ApApCpApA product with no visible n–1 or initial primer present. The yield of final product was about 90% of the initial primer.

EXAMPLE 5

Synthesis of dApdA

The oligonucleotide dApdA was synthesized by initially synthesizing the oligonucleotide (ApApC)-pdApdA using the initial primer ApApC and two coupling cycles with the activated nucleotide AppdAp. Synthesized oligodeoxyribonucleotide dApdA was cleaved from the initial primer using RNase treatment.

The following solution was placed in a total volume of 30 ul in an ependorf tube: 50 mM Tris-Cl, pH 8.0, 10 mM $MgCl_2$, 10 mM DTT, 0.1% TRITON X-100, 1 mM ApApC primer, 5 mM AppdAp. The following procedures were performed:

cycle 1

(a) Add 1 ul (20 units) RNA ligase (phage T4, New England Biolabs). Incubate at 37° C. for 3 hours. Heat at 95° C. for 5 minutes, cool to room temperature.

(b) Add 1 ul (0.03 units) Nucleotide Pyrophosphatase (snake venom, Sigma P7383). Incubate at 37° C. for 30 minutes. Heat at 95° C. for 5 minutes, cool to room temperature.

(c) Add 1 ul (3 units) Alkaline Phosphatase (calf intestine, US Biochemicals). Incubate at 37° C. for 30 minutes. Heat at 95° C. for 5 minutes, cool to room temperature.

cycle 2

(a) Add 10 ul 10 mM AppdAp+1 ul RNA ligase. Incubate at 37° C. for 20 hours. Heat at 95° C. for 5 minutes, cool to room temperature.

(b) same as cycle 1

(c) same as cycle 1

Insoluble coagulated protein debris was removed by pelleting at 12,000 g for 5 min. USB TLC revealed pure ApApCpdApdA product with no visible n–1 or initial primer present. The yield of final product was about 90% of the initial primer. Cleavage of the synthesized oligodeoxyribonucleotide dApdA from the oligonucleotide product was performed by adding 100 ng RNase A (bovine pancreas, US Biochemicals) to 4 ul oligonucleotide product and incubating at 37° C. for 1 hour. dApdA product was analyzed and purified from nucleosides and ApApCp using butyric TLC.

EXAMPLE 6

Synthesis of (ApApC)-pApA

The following solution was placed in a total volume of 30 ul in an ependorf tube: 50 mM Tris-Cl, pH 8.0, 10 mM $MgCl_2$, 10 mM DTT, 0.1% TRITON X-100, 1 mM ApApC primer, 5 mM AppAp containing 10% glycerol as a preservative. The solution was overlaid with 50 ul light mineral oil to prevent evaporation. The following procedures were performed:

cycle 1

(a) Add 1 ul (20 units) RNA ligase (phage T4, New England Biolabs)+0.5 ul (0.2 units) Inorganic Pyrophosphatase (Sigma, yeast)+0.5 ul (0.025 units) 5'-Nucleotidase (Sigma, snake venom). Incubate at 37° C. for 1 hour. Heat at 95° C. for 5 minutes, cool to room temperature.

(b) Add 1 ul (0.03 units) Nucleotide Pyrophosphatase (Sigma P7383, snake venom). Incubate at 37° C. for 30 minutes. Heat at 95° C. for 5 minutes, cool to room temperature.

(c) Add 1 ul (3 units) Alkaline Phosphatase (calf intestine, US Biochemicals)+0.5 ul (0.05 units) Nucleoside Phosphorylase (Sigma). Incubate at 37° C. for 30 minutes. Heat at 95° C. for 5 minutes, cool to room temperature.

cycle 2

(a) Add 10 ul 10 mM AppAp+1 ul (20 units) RNA ligase. Incubate at 37° C. for 5.5 hours. Heat at 95° C. for 5 minutes, cool to room temperature.

(b) same as cycle 1

(c) same as cycle 1

Insoluble coagulated protein debris was removed by adding 5 ug proteinase K (Sigma) and incubating at 60° C. for 5 minutes. This treatment removed most of the debris. The proteinase K was heat inactivated at 95° C. for 5 minutes, then cooled to room temperature. Mineral oil was removed with a pipettor. Residual mineral oil was removed by adding 100 ul chloroform, vortexed vigorously, and centrifuged at 12,000 g for 1 minute to separate the phases. The chloroform extraction also removed protein from the aqueous phase, which appeared between the two phases. The upper aqueous phase was collected by pipettor and was analyzed by USB TLC. This revealed pure ApApCpApA product with no visible n–1 or initial primer present. The yield of final product was about 90% of the initial primer.

EXAMPLE 7

Synthesis of (ApApC)-pApA

The following solution was placed in a total volume of 30 ul in an ependorf tube: 50 mM Tris-Cl, pH 8.0, 10 mM $MgCl_2$, 10 mM DTT, 0.1% TRITON X-100, 1 mM ApApC primer, 5 mM AppAp. The following procedures were performed:

cycle 1.

(a) Add 1 ul (20 units) RNA ligase (phage T4, New England Biolabs). Incubate at 37° C. for 1 hour. Add 1 ug Proteinase K (Sigma), incubate at 60° C. for 5 minutes, heat at 95° C. for 5 minutes to inactivate protease, and cool to room temperature.

(b) Add 1 ul (0.03 units) Nucleotide Pyrophosphatase (snake venom, Sigma P7383). Incubate at 37° C. for 30 minutes. Add 1 ug Proteinase K (Sigma), incubate at 60° C. for 5 minutes, heat at 95° C. for 5 minutes to inactivate protease, and cool to room temperature.

(c) Add 1 ul (3 units) Alkaline Phosphatase (calf intestine, US Biochemicals). Incubate at 37° C. for 30 minutes. Add 1 ug Proteinase K (Sigma), incubate at 60° C. for 5 minutes, heat at 95° C. for 5 minutes to inactivate protease, and cool to room temperature.

cycle 2

(a) Add 10 ul 10 mM AppAp+1 ul (20 units) RNA ligase. Incubate at 37° C. for 5.5 hours. Add 1 ug Proteinase K (Sigma), incubate at 60° C. for 5 minutes, heat at 95° C. for 5 minutes to inactivate protease, and cool to room temperature. (b) same as cycle 1

(c) Add 1 ul (3 units) Alkaline Phosphatase (calf intestine, US Biochemicals). Incubate at 37° C. for 30 minutes. The use of Proteinase K for the inactivation of enzymes after each step prevented the accumulation of insoluble coagulated protein debris. USB TLC revealed pure ApApC-pApA product with no visible n–1 or initial primer present. The yield of final product was about 90% of the initial primer.

EXAMPLE 8

Synthesis of (ApApC)-pApA

The following solution was placed in a total volume of 30 ul in an ependorf tube: 50 mM Tris-Cl, pH 8.0, 10 mM $MgCl_2$, 10 mM DTT, 0.1% TRITON X-100, 1 mM ApApC primer, 5 mM AppAp, containing 10% dimethylsulfoxide to inhibit base pairing. The synthesis procedure was identical to Example 3. USB TLC revealed pure-ApApCpApA product with no visible n–1 or initial primer present. The yield of final product was about 90% of the initial primer.

EXAMPLE 9

Synthesis of (ApApC)-pApA

The following solution was placed in a total volume of 30 ul in an ependorf tube: 50 mM Tris-Cl, pH 8.0, 10 mM $MgCl_2$, 10 mM DTT, 0.1% TRITON X-100, 1 mM ApApC primer, 5 mM AppAp, 10 uM Vanadyl Ribonucleoside Complexes (to inhibit any contaminating RNases). The synthesis procedure was identical to Example 3. USB TLC revealed pure ApApCpApA product with no visible n–1 or initial primer present. The yield of final product was about 90% of the initial primer.

EXAMPLE 10

Synthesis of (ApApC)-pApA

The following solution was placed in a total volume of 30 ul in an ependorf tube: 50 mM Tris-Cl, pH 8.0, 10 mM $MgCl_2$, 10 mM DTT, 0.1% TRITON X-100, 1 mM ApApC primer, and 5 mM AppAp. The following procedures were performed:

cycle 1

(a) Add 1 ul (20 units) RNA ligase (phage T4, New England Biolabs)+1 ul 3 mM sodium pyrophosphate+1 ul 300 mM glucose+0.2 units hexokinase (yeast, Sigma). Incubate at 37° C. for 1 hour. Heat at 95° C. for 5 minutes, cool to room temperature.

(b) Add 1 ul (0.03 units) Nucleotide Pyrophosphatase (snake venom, Sigma P7383). Incubate at 37° C. for 30 minutes. Heat at 95° C. for 5 minutes, cool to room temperature.

(c) Add 1 ul (3 units) Alkaline Phosphatase (calf intestine, US Biochemicals). Incubate at 37° C. for 30 minutes. Heat at 95° C. for 5 minutes, cool to room temperature.

cycle 2

(a) Add 10 ul 10 mM AppAp+1 ul (20 units) RNA ligase+1 ul 3 mM sodium pyrophosphate+1 ul 300 mM glucose+0.2 units hexokinase. Incubate at 37° C. for 3.5 hours. Heat at 95° C. for 5 minutes, cool to room temperature.

(b) same as cycle 1

(c) same as cycle 1

Insoluble coagulated protein debris was removed by pelleting at 12,000 g for 5 min. USB TLC revealed nearly pure ApApCpApA product with slight n–1 side product.

EXAMPLE 11

One Pot Synthesis of ApApC-pApA with TAP

The following solution was placed in a total volume of 30 ul in an ependorf tube: 50 mM BES, pH 7.0, 10 mM $MgCl_2$, 10 mM DTT, 0.1% TRITON X-100, 1 mM ApApC primer, 5 mM AppAp. The following procedures were performed:

cycle 1

(a) Add 2 ul (40 units) RNA Ligase (phage T4, New England Biolabs). Incubate at 37° C. for 2 hours. Heat at 95° C. for 5 minutes, cool to room temperature.

(b) Add 1 ul (3 units) Alkaline Phosphatase (calf intestine, US Biochemicals)+1 ul (2 units) Tobacco Acid Pyrophosphatase (Sigma). Incubate at 37° C. for 3.5 hours. Heat at 95° C. for 5 minutes, cool to room temperature.

cycle 2

(a) Add 10 ul 10 mM AppAp+1 ul RNA Ligase. Incubate at 37° C. for 30 minutes. Heat at 95° C. for 5 minutes, cool to room temperature.

(b) same as cycle 1

Insoluble coagulated protein debris was removed by pelleting at 12,000 g for 1 min. The only oligonucleotide product visible by USB TLC was the desired oligonucleotide product ApApCpApA. The final yield of oligonucleotide product was nearly 100%.

EXAMPLE 12

Synthesis of ApApC-pdApdA Using TdT+ddATP Capping

The following solution was placed in a total volume of 30 ul in an ependorf tube: 50 mM Tris-Cl, pH 8.0, 10 mM $MgCl_2$, 10 mM DTT, 0.1% TRITON X-100, 1 mM ApApC primer, 5 mM AppdAp. The following procedures were performed:

cycle 1

(a) Add 1 ul (20 units) RNA Ligase (phage T4, New England Biolabs). Incubate at 37° C. for 3 hours. Heat at 95° C. for 5 minutes, cool to room temperature.

(b) Add 1 ul Terminal deoxynucleotidyl Transferase (USB, 17 units/ul)+3 ul 5 mM dideoxyadenosine 5'-triphosphate. Incubate at 37° C. for 2.5 hours. Add 1 ug Proteinase K, incubate at 60° C. for 15 minutes, heat at 95° C. for 5 min, cool to room temperature.

(c) Add 1 ul (0.03 units) Nucleotide Pyrophosphatase (snake venom, Sigma P7383). Incubate at 37° C. for 30 minutes. Heat at 95° C. for 5 minutes, cool to room temperature.

(d) Add 1 ul (3 units) Alkaline Phosphatase (calf intestine, US Biochemicals). Incubate at 37° C. for 30 minutes. Heat at 95° C. for 5 minutes, cool to room temperature.

cycle 2

(a) Add 10 ul 10 mM AppdAp+1 ul RNA Ligase. Incubate at 37° C. for 15 hours. Heat at 95° C. for 5 minutes, cool to room temperature.

(b) same as cycle 1

(c) same as cycle 1

(d) same as cycle 1

Insoluble coagulated protein debris was removed by pelleting at 12,000 g for 5 min. The reaction mixture supernatant was analyzed by USB TLC. The only oligonucleotide prod-

uct visible was the desired product ApApCpdApdA. The yield of final product was about 90% of the initial primer.

EXAMPLE 13

Synthesis of ApApC-DApA Using Enzyme-Solid Support Matrix

The following solution was placed in a total volume of 30 ul in an ependorf tube: 50 mM Tris-Cl, pH 8.0, 10 mM $MgCl_2$, 10 mM DTT, 0.1% TRITON X-100, 1 mM ApApC primer, 5 mM AppAp. The following procedures were performed:

cycle 1

(a) Add 1 ul (20 units) RNA Ligase (phage T4, New England Biolabs). Incubate at 37° C. for 1 hour. Heat at 95° C. for 5 minutes, cool to room temperature.

(b) Add 1 ul (0.03 units) Nucleotide Pyrophosphatase (snake venom, Sigma P7383). Incubate at 37° C. for 30 minutes. Heat at 95° C. for 5 minutes, cool to room temperature.

(c) Add 6 ul Alkaline Phosphatase-Acrylic Beads (calf intestine, Sigma Chemical Co.). Incubate at 37° C. for 2.5 hours with occasional mixing. Remove CIAP-acrylic beads by briefly pelleting. Heat supernatant at 95° C. for 5 minutes to remove any residual CIAP leakage, cool to room temperature.

cycle 2

(a) Add 10 ul 10 mM AppAp+1 ul RNA Ligase. Incubate at 37° C. for 2 hours. Heat at 95° C. for 5 minutes, cool to room temperature.

(b) same as cycle 1

(c) same as cycle 1, except skip the heat inactivation

Insoluble coagulated protein debris was removed by pelleting at 12,000 g for 5 min. USB TLC revealed a mixture of approximately 50% ApApCpA and 50% ApApCpApA oligonucleotide product. The n−1 failure sequence was due to the incomplete 3'-dephosphorylation of the oligonucleotide in the first cycle. This example demonstrates that the enzymes can be covalently attached to a solid matrix.

EXAMPLE 14

The method of the invention can be used for synthesizing oligonucleotide mixtures in which two or more different bases are used at a particular position. This technique is known in the art as "wobbling" and is useful in hybridization applications of an oligonucleotide to a DNA library based on amino acid sequence. Wobbling is performed by adding a mixture of blocked nucleotide substrates instead of a single blocked nucleotide substrate during the RNA ligase step of one cycle. The relative amounts of the blocked nucleotides used is selected to balance out differences in coupling rate. For example, if a 50:50 mix of A and G is desired, a mixture of the nucleotide substrates AppAp and AppGp would be added during the RNA ligase step of the appropriate reaction cycle. If the reactivities of AppAp and AppGp are equal, the substrates would be used in equal amounts.

EXAMPLE 15

Synthesis of (ApApC)-pApA

The same protocol was used from example 3, except that after RNA Ligase coupling, the heat inactivation step in part (a) of each cycle was omitted, and 20 units additional RNA Ligase was added during each Alkaline Phosphatase digestion. USB TLC revealed pure ApApCpApA product with no visible n−1 or initial primer present. The yield of final product was about 90% of the initial primer.

This example demonstrates that inactivation of the chain extending enzyme is not necessary. In addition, the use of a thermostable chain extending enzyme would obviate the need to add this enzyme after each cycle. This example also demonstrates that phosphodiesterase I incubation can be performed without prior inactivation of the chain extending enzyme. Optionally, phosphodiesterase I incubation can be performed in the presence of 5'-Nucleotidase to hydrolyze AMF generated by phosphodiesterase I cleavage of App-(d)Np.

EXAMPLE 16

Synthesis of ApApCpApA with Substrate Reuse

The oligonucleotide ApApCpApA was synthesized according to the following procedure. The following solution was placed in a total volume of 30 ul in an ependorf tube: 50 mM Tris-Cl, pH 8.0, 10 mM $MgCl_2$, 10 mM DTT, 0.1% TRITON X-100, 1 mM ApApC initial primer, and Nucleotide Substrate. The following procedure was performed:

cycle 1

(a) Add 1 ul (20 units) T4 RNA Ligase (New England Biolabs), incubate at 37 degrees C. for 3 hours, heat at 85 degrees C. for minutes, cool.

(b) Add 1 ul (3 units) T4 Polynucleotide Kinase (US Biochemicals, contains 3'-Phosphatase), incubate at 37 degrees C. for 1 hour, heat at 85 degrees C for 5 minutes, cool.

cycle 2 - starting volume is 20 ul

(a) same as cycle 1. No AppAp substrate was added.

(b) same as cycle 1.

Sub-Example A: Nucleotide substrate was approximately 5 mM AppAp.

Sub-Example B: Nucleotide substrate was 5 mM 3',5'ADP+ 4.5 mM ATP. These precursors are converted to AppAp in the first cycle by RNA Ligase. Supplementation with inorganic pyrophosphatase in a separate experiment improved oligonucleotide product yield.

USB TLC confirmed the formation of ApApCpApA product for both sub-examples. USB TLC also confirmed that no significant inactivated nucleotide substrate AppA was formed for both sub-examples. Approximately 5 ul oligonucleotide product was incubated with 100 ng RNase A (US Biochemicals) at 37° C. for about 15 minutes. RNase A is used as a base-specific RNase to cleave the oligonucleotide 3' to the Cytidine base. Butyric TLC confirmed the formation of ApA oligonucleotide product for both sub-examples. Yield of oligonucleotide product was better in sub-example A.

This experiment demonstrates reuse in the Second cycle of nucleotide substrate AppAp used in the first cycle. This was accomplished by using bacteriophage T4 3'-Phosphatase under carefully controlled conditions to specifically remove the extended primer blocking group without significantly inactivating the nucleotide substrate AppAp. The high concentration of primer and nucleotide substrate used in this example and the following examples is for the convenience of allowing detection of product by thin layer chromatography. Proportionately lower concentrations, such as 0.10 mM primer and 1.0 mM nucleotide substrate may be more appropriate for long oligonucleotides to lessen the build up of side products.

EXAMPLE 17

Synthesis of ApApCpApA using Rye Grass 3'-Phosphatase

ApApCpApA was synthesized using the same procedure as Example 16, sub-example A, except 0.05 units 3'-Phosphatase from Rye Grass (Sigma, sold as 3'-Nucleotidase) was used for 3 hours at 37 degrees C. in place of T4 Polynucleotide Kinase (3'-Phosphatase). Butyric TLC confirmed synthesis of product and RNase A digestion confirmed formation of ApA.

EXAMPLE 18

Synthesis of ApApCpApA using Preferred Mode with Substrate Reuse

The following solution was placed in a total volume of 30 ul in an ependorf tube: 50 mM Tris-Cl, pH 8.0, 10 mM $MgCl_2$, 10 mM DTT, 0.1% Triton X-100, 1 mM ApApC initial primer, and 5 mM AppAp. The following procedure was performed:

cycle 1

(a) Add 1 ul (20 units) T4 RNA Ligase (New England Biolabs)+0.5 ul (0.025 units) 5'-Nucleotidase (Sigma), incubate at 37 degrees C. for 1 hour, heat at 85 degrees C. for 5 minutes, cool.

(b) Add Exonuclease—see details below. Heat at 95° C. for 5 minutes, cool.

(c) Add 0.5 ul (15 units) T4 Polynucleotide Kinase (US Biochemicals), incubate at 37 degrees C for 30 minutes, heat at 85 degrees C. for 5 minutes, cool.

cycle 2—starting volume is 20 ul

(a) same as cycle 1, but incubation is extended to 135 minutes. No AppAp substrate was added.

(b) same as cycle 1.

(c) same as cycle 1.

Sub-Example A: Exonuclease added was 1 ul (0.02 units) Phosphodiesterase I (US Biochemicals). In this sub-example only, 1 ul 100 mM ATP is added during RNA Ligase incubation in the second cycle to reform the substrate AppAp from 3',5'-ADP.

Sub-Example B: Exonuclease added was 1 ul (10 units) Exonuclease I (US Biochemicals)

Sub-Example C: Exonuclease added was 1 ul (0.1 units) Polynucleotide Phosphorylase (Sigma). In this sub-example only, 0.2 mM Na_2AsO_4 was incorporated in the buffer through-out the synthesis to facilitate Polynucleotide Phosphorylase digestion of unextended primer chains.

USB TLC confirmed the formation of ApApCpApA product in all sub-examples. Digestion with RNase A confirmed the formation of ApA in all sub-examples.

EXAMPLE 19

Synthesis of ApApCpApA using Substrate Reuse, and AMP Inactivating Enzyme(s)

The following solution was placed in a total volume of 30 ul in an ependorf tube: 50 mM Tris-Cl, pH 8.0, 10 mM $MgCl_2$, 10 mM DTT, 0.1% Triton X-100, 1 mM ApApC initial primer, and 5 mM AppAp. The following procedure was performed:

cycle 1

(a) Add 1 ul (20 units) T4 RNA Ligase (New England Biolabs)+AMP Inactivating Enzyme(s), incubate at 37° C. for 3 hours, heat at 85° C. for 5 minutes, cool.

(b) Add 1 ul (3 units) T4 Polynucleotide Kinase (US Biochemical), incubate at 37° C. for 1 hour, heat at 85° C. for 5 minutes, cool.

cycle 2

(a) same as cycle 1. No AppAp substrate is added.

(b) same as cycle 1.

Sub-Example A: AMP Inactivating Enzyme was 0.5 ul (0.025 units) 5'-Nucleotidase (Sigma)

Sub-Example B: AMP Inactivating Enzyme was 0.5 ul (0.025 units) 5'-Nucleotidase (Sigma)+1 ul (0.018 units) Adenosine Deaminase (Sigma).

Sub-Example C: AMP Inactivating Enzyme was 1 ul (0.004 units) AMP Deaminase (Sigma).

Sub-Example D: AMP Inactivating Enzyme was 1 ul (0.12 units) AMP Nucleosidase (E. coli).

USB TLC confirmed the formation of ApApCpApA product in all sub-examples. USB TLC also confirmed that the AMP Inactivating Enzymes in all sub-examples converted substantially all substrate to product. In all sub-examples, butyric TLC confirmed that the oligonucleotide ApA was cleaved from the product by RNase A digestion. It was also found that adenosine deaminase was not inactivated at 95° C., a useful property.

EXAMPLE 20

Synthesis of ApApCpApApdA using cycles with and without Substrate Reuse

The following solution was placed in a total volume of 30 ul in an ependorf tube: 50 mM Tris-Cl, pH 8.0, 10 mM $MgCl_2$, 10 mM DTT, 0.1% Triton X-100, 1 mM ApApC initial primer, and 5 mM AppAp. The following procedure was performed:

cycle 1: Reuse

(a) add 1 ul (20 units) T4 RNA Ligase (New England Biolabs), incubate at 37 degrees C. for 1 hour, heat at 85° C. for 5 minutes, cool.

(b) add 1 ul (3 units) T4 Polynucleotide Kinase (US Biochemicals), incubate at 37 degrees C. for 1 hour, heat at 85° C. for 5 minutes, cool.

cycle 2: No Reuse

(a) add 1 ul (20 units) T4 RNA Ligase (New England Biolabs), incubate at 37 degrees C. for 1 hour, heat at 85° C. for 5 minutes, cool.

(b) add 1 ul (0.035 units) Nucleotide Pyrophosphatase (Sigma, snake venom), incubate at 37° C. for 30 minutes, heat at 95° C. for 5 minutes, cool.

(c) add 1 ul (1.6 units) Alkaline Phosphatase (US Biochemicals, calf intestine), incubate at 45° C. for 30 minutes, heat at 95° C. for 5 minutes, cool. (Alkaline Phosphatases generally have better activity at higher temperatures, such as 45°–60° C.).

cycle 3: No Reuse

(a) add 2 ul (40 units) T4 RNA Ligase (New England Biolabs)+10 ul 10 mM AppdAp, incubate at 37° C. for 80 minutes, heat at 85° C. for 5 minutes, cool.

(b) same as cycle: 2.

(c) same as cycle 2.

USB TLC strongly suggested formation of ApApCpApApdA product. Incubation of 5 ul oligonucleotide product with 100 ng RNase A (US Biochemicals) at 37° C. for 15 minutes resulted in the cleavage of the oligonucleotide to ApApdA product as strongly suggested by USB and buytric. Matrix assisted laser desorption mass spectroscopy confirms formation of this product.

I claim:

1. A method for synthesizing an oligonucleotide of a defined sequence, comprising the steps of:

(a) combining (1) an oligonucleotide primer and (2) a blocked nucleotide or a blocked nucleotide precursor that forms a blocked nucleotide in situ, in a reaction mixture in the presence of a chain extending enzyme effective to couple the blocked nucleotide to the 3'-end of the oligonucleotide primer such that a primer-blocked nucleotide product is formed, wherein the blocked nucleotide comprises a nucleotide to be added to form part of the defined sequence and a blocking group attached to the 3'-end of the nucleotide effective to prevent the addition of more than one blocked nucleotide to the primer;

(b) removing the blocking group from the 3'-end of the primer-blocked nucleotide product to form a primer-nucleotide product, whereby the reaction mixture contains any unreacted starting materials that may remain, primer-nucleotide product and reaction by-products; and

(c) repeating at least one cycle of steps (a) and (b) using the primer-nucleotide product from step (b) as the oligonucleotide primer of step (a) in the subsequent cycle without separation of the primer-nucleotide product from the remainder of the reaction mixture.

2. A method according to claim 1, wherein the blocking group is removed enzymatically.

3. A method according to claim 2, wherein each cycle further comprises the additional step of inactivating unreacted blocked nucleotide in the reaction mixture to render it less reactive as a substrate for chain extending enzyme.

4. A method according to claim 3, wherein the chain extending enzyme is RNA ligase.

5. A method according to claim 4, wherein the blocked nucleotide is App(d)Np, where N represents any nucleoside or nucleoside analog which RNA ligase can couple to an oligonucleotide primer.

6. A method according to claim 5, wherein the blocking group is a phosphate and is removed from the primer-blocked nucleotide product by a phosphatase.

7. A method according to claim 5, wherein unreacted blocked nucleotide is inactivated by a phosphatase enzyme.

8. A method according to claim 5, wherein the unreacted blocked nucleotide is inactivated by a Dinucleotide Pyrophosphate Degrading Enzyme.

9. A method according to claim 1, wherein each cycle of steps further comprises the additional step of modifying uncoupled oligonucleotide primer to prevent its coupling to blocked nucleotide in subsequent cycles of the method.

10. A method according to claim 9, wherein the uncoupled oligonucleotide primer is modified by incubating with at least one Exonuclease, whereby the uncoupled oligonucleotide primer is degraded.

11. A method according to claim 9, wherein the uncoupled oligonucleotide primer is modified by incubating with a chain terminating nucleotide and an enzyme effective to couple the chain terminating nucleotide to uncoupled oligonucleotide primer, whereby uncoupled oligonucleotide primer is terminated.

12. A method according to claim 2, wherein the defined sequence includes at least one repeat region which is synthesized by a method comprising the steps of:

(1) extending the oligonucleotide primer with 3'-phosphate-blocked nucleotide to form 3'-phosphate-blocked primer-nucleotide;

(2) removing the 3'-phosphate blocking group from the 3'-phosphate-blocked primer-nucleotide substantially without removing the 3'-phosphate blocking group from unreacted 3'-phosphate-blocked nucleotide using 3'Phosphatase; and

(3) repeating steps (1) and (2) using unreacted 3'-phosphate-blocked nucleotide from step (2) as the 3'-phosphate-blocked nucleotide of step (1).

13. A method for synthesizing a repeat region of an oligonucleotide having a defined sequence, said repeat region including a repeated nucleotide that appears more than once in succession, comprising the steps of:

(a) enzymatically coupling an oligonucleotide primer with a 3'-phosphate-blocked repeated nucleotide to form a 3'-phosphate blocked primer-nucleotide;

(b) removing the 3'-phosphate blocking group from the 3'-phosphate-blocked primer-nucleotide using 3'-phosphatase enzyme substantially without removing the 3'-phosphate blocking group from unreacted 3'-phosphate-blocked repeated nucleotide; and

(c) repeating steps (a) and (b) using the unreacted 3'-phosphate-blocked repeated nucleotide from step (b) as the 3'-phosphate-blocked nucleotide of step (a) and using the deblocked primer-nucleotide product of step (b) as the oligonucleotide primer of step (a).

14. A method for synthesizing an oligonucleotide, wherein the 3'-end of an oligonucleotide primer is coupled with a blocked nucleotide to form a primer-blocked nucleotide product in a reaction mixture, said blocked nucleotide comprising a nucleotide to be added to the oligonucleotide primer and a blocking group attached to the 3'-end of the nucleotide effective to prevent the addition of more than one blocked nucleotide to the oligonucleotide primer, comprising incubating the reaction mixture with an exonuclease, whereby any oligonucleotide primer which was not coupled is degraded, substantially without degrading the primer-blocked nucleotide product.

15. A method for synthesizing an oligonucleotide, wherein the 3'-end of an oligonucleotide primer is enzymatically coupled with a blocked nucleotide to form a primer-blocked nucleotide product in a reaction mixture, said blocked nucleotide comprising a nucleotide to be added to the oligonucleotide primer and a removable blocking group attached to the 3'-end of the nucleotide effective to prevent the addition of more than one blocked nucleotide to the oligonucleotide primer, comprising incubating the reaction mixture with a chain terminating nucleotide and an enzyme effective to couple the chain terminating nucleotide to the oligonucleotide primer, whereby oligonucleotide primer which was not coupled to a blocked nucleotide is end-capped to render it unreactive to further coupling, said chain terminating nucleotide being different from said blocked nucleotide and selected such that end-capped oligonucleotide primer remains end-capped and unreactive when the blocking group is removed from the primer-blocked nucleotide product.

16. A method according to claim 15, wherein the chain terminating nucleotide is a dideoxynucleotide.

17. A method according to claim 14, wherein at least two nucleotides are added to the primer without intermediate purification of the resulting oligonucleotide product from other reactants and reaction by-products.

18. A method according to claim 1, wherein each cycle further comprises the additional step of converting adenosine monophosphate released in the coupling reaction to a less reactive form, whereby any inhibitory effect of the adenosine monophosphate pm the coupling of the oligonucleotide primer to the blocked nucleotide is minimized.

19. A method according to claim 1, further comprising the step of cleaving a synthesized oligonucleotide to remove

some or all of the oligonucleotide primer used in the first cycle of the method from the synthesized oligonucleotide.

20. A method for coupling a blocked nucleotide AppNp to an oligonucleotide primer, characterized in that the blocked nucleotide is coupled to the primer using RNA Ligase in the absence of ATP, and in that pyrophosphate or unactivated nucleotide substrate, 3,5'-NDP, is used to regenerate free RNA Ligase from the inactivated adenylylated form, wherein N represents any nucleoside or nucleoside analog which RNA ligase can couple to an oligonucleotide primer.

21. A method for coupling a blocked nucleotide to an oligonucleotide primer, characterized in that the coupling is performed using RNA Ligase in the presence of 5'-Nucleotidase, AMP Nucleotidase or AMP Deaminase, whereby AMP released in the coupling reaction is converted to a form which is less effective than AMP to inhibit the coupling reaction or participate as a substrate in a reverse coupling reaction.

22. A method for synthesizing a selected oligonucleotide wherein an oligonucleotide primer is extended by enzymatically adding at least two nucleotides to the 3'-end of the oligonucleotide primer, characterized in that the primer is cleaved from the added nucleotides to form the selected oligonucleotide.

23. A method for converting a blocked nucleotide comprising a dinucleotide pyrophosphate moiety, a blocking group effective to prevent the enzymatic coupling of more than one blocked nucleotide to an oligonucleotide primer, and a nucleotide to be enzymatically coupled to the primer to a less reactive form, characterized in that the blocked nucleotide is treated with a dinucleotide pyrophosphate degrading enzyme.

24. A method according to claim **8**, wherein the Dinucleotide Pyrophosphate Degrading Enzyme is Nucleotide Pyrophosphatase.

25. A method according to claim **23**, wherein the Dinucleotide Pyrophosphate Degrading Enzyme is Nucleotide Pyrophosphatase.

26. A method according to claim **10**, wherein the Exonuclease is exonuclease I, phosphodiesterase I, or polynucleotide kinase.

27. A method according to claim **14**, wherein the Exonuclease is exonuclease L phosphodiesterase I, or polynucleotide kinase.

28. A method according to claim **18**, wherein adenosine monophosphate is converted to a less reactive form using 5'-Nucleotidase.

29. A method according to claim **18**, wherein adenosine monophosphate is converted to a less reactive form using AMP Nucleosidase.

* * * * *

RELATED PROCEEDING APPENDIX

None